统计学教程

第四军医大学

徐勇勇

第一章　绪 论

一、教学大纲要求

（一）掌握内容

1．几个基本概念

样本与总体、频率与概率、资料类型、随机变量、误差。

2．统计工作的步骤

设计、收集资料、整理资料、分析资料。

（二）熟悉内容

医学统计学的含义、内容及其医学应用。

（三）了解内容

医学统计的历史发展。

二、教学内容精要

(一) 统计学、医学统计学、卫生统计学

统计学是研究数据的收集、整理、分析与推断的科学。

医学统计学是用统计学的原理和方法研究生物医学现象的一门学科。

卫生统计学则是把统计理论、方法应用于居民健康状况研究、医疗卫生实践、卫生事业管理和医学科研的一门应用学科。

(二) 统计学中的几个基本概念

1．随机变量

随机变量（random variable**）指取值不能事先确定的观察结果，通常简称为变量。随机变量有一个共同的特点是不能用一个常数来表示，而且理论上讲，每个变量的取值服从特定的概率分布。**

随机变量可分为两种类型：离散型变量和连续型变量。

2．误差

误差（error**）指实际观察值与观察真值之差、样本指标与总体指标之差。误差可分为系统误差和随机误差，两种误差的区别见表** 1-1**。**

表 1-1　**系统误差与随机误差的区别**

误差分类	产生原因	对观察值的影响	处理方法
系统误差	仪器未校正、测量者感官的某种偏差、医生掌握疗效标准偏高或偏低等。	使观察值不是分散在真值的两侧，而是有方向性、系统性或周期性地偏离真值。	通过实验设计的完善和技术措施的改进来消除或减少。
随机误差	排除系统误差后，其他多种不确定因素。	使观察值不按方向性、系统性而随机的变化，误差变量	可通过统计处理估计随机误差。

一般服从正态分布。

3．资料类型

观察单位的某项特征的测量结果按其性质可分为三种类型：

（1）计量资料：对每个观察单位用定量的方法测定某项指标量的大小，所得的资料称为计量资料（measurement data）。计量资料亦称定量资料、测量资料。其变量值是定量的，表现为数值大小，一般有度量衡单位。如某一患者的身高（cm）、体重(kg)、红细胞计数($10^{12}/L$)、脉搏（次/分）、血压（KPa）等。

（2）计数资料：将观察单位按某种属性或类别分组，所得的观察单位数称为计数资料（count data）。计数资料亦称定性资料或分类资料。其观察值是定性的，表现为互不相容的类别或属性。如调查某地某时的男、女性人口数；治疗一批患者，其治疗效果为有效、无效的人数；调查一批少数民族居民的A、B、AB、O 四种血型的人数等。

（3）等级资料：将观察单位按测量结果的某种属性的不同程度分组，所得各组的观察单位数，称为等级资料（ordinal data）。等级资料又称有序变量。如患者的治疗结果可分为治愈、好转、有效、无效或死亡，各种结果既是分类结果，又有顺序和等级差别，但这种差别却不能准确测量；一批肾病患者尿蛋白含量的测定结果分为 +、++、+++等。

等级资料与计数资料不同：属性分组有程度差别，各组按大小顺序排列。

等级资料与计量资料不同：每个观察单位未确切定量，故亦称为半计量资料。

4．总体与样本

总体（population）指特定研究对象中所有观察单位的测量值。可分为有限总体和无限总体。总体中的所有单位都能够标识者为有限总体，反之为无限总体。

从总体中随机抽取部分观察单位，其测量结果的集合称为样本（sample）。样本应具有代表性。所谓有代表性的样本，是指用随机抽样方法获得的样本。

5．概率与频率

概率（probability）又称几率，是度量某一随机事件 A 发生可能性大小的一个数值，记为 $P（A）$。$0<P（A）<1$。

在相同的条件下，独立重复做 n 次试验，事件 A 出现了 m 次，则比值 m/n 称为随机事件 A 在 n 次试验中出现的频率（frequency）。当试验重复很多次时 $P（A）= m/n$。

（三）统计工作的步骤

1．设计：设计内容包括资料收集、整理和分析全过程总的设想和安排。设计是整个研究中最关键的一环，是今后工作应遵循的依据。

2．收集资料：应采取措施使能取得准确可靠的原始数据。

3．整理资料：简化数据，使其系统化、条理化，便于进一步分析计算。

4．分析资料：计算有关指标，反映事物的综合特征，阐明事物的内在联系和规律。分析资料包括统计描述和统计推断。

（四）医学研究中统计方法的应用

医学统计方法在医学研究中的应用主要有三个方面：

1．以正确的方式收集数据；

2．描述数据的统计特征；

3．统计分析得出正确结论。

（五）医学统计历史发展

最初的统计主要是数据汇总。统计发展到今天，已经成为一种对客观现象数量方面进行的调查研究活动，是收集、整理、分析、判断等认识活动的总称，数据汇总仅仅是统计工作的一小部分。医学统计的发展起源于生物统计、遗传统计，经过了描述统计、大样本统计、小样本统计推断、临床统计和多变量统计几个发展阶段。

三、典型试题分析

（一） 名词解释

抽样误差。

答案：抽样误差（sampling error）是指样本统计量与总体参数的差别。在总体确定的情况下，总体参数是固定的常数，统计量是在总体参数附近波动的随机变量。

[评析] 本题考点：抽样误差的概念。

抽样误差是统计学中的重要概念。在抽样研究中是不可避免的。产生抽样误差的根本原因是生物个体间存在的变异性。

（二）单项选择题

1．统计学中所说的样本是指（ ）。

A．随意抽取的总体中任意部分

B．有意识的选择总体中的典型部分

C．依照研究者要求选取总体中有意义的一部分

D．依照随机原则抽取总体中有代表性的一部分

答案：D

[评析] 本题考点：统计学中样本概念的理解。

统计学中的样本是指从总体中随机抽取的部分观察单位测量值的集合。这里的“随机抽取”并非通常所说的“随意抽取”，而是保证总体中每个观察单位等概率被抽取的科学方法。随机抽样是样本具有代表性的保证。

2．下列资料属等级资料的是（ ）。

A．白细胞计数　　B．住院天数

C．门急诊就诊人数　　D．病人的病情分级

答案： D

[评析] 本题考点：统计资料的分类。

统计资料按其性质可分为三种类型：计量资料、计数资料和等级资料。计量资料变量值是定量的，表现为数值大小，一般有度量衡单位，如本例中白细胞计数（$10^9/L$），住院天数（天）。计数资料其观察值是定性的，表现为互不相容的类别或属性的观察单位数，如门急诊就诊人数可按门诊、急诊分类清点各组人数。等级资料的属性分组有程度差别，各组按大小顺序排列，如病人的病情分级为轻、中、重。

（三）简答题

一位研究人员欲做一项实验研究，研究设计应包括那几方面的内容？

答案：一般来讲，研究设计应包括两方面的设计：专业设计和统计设计。专业设计是针

对专业问题进行的研究设计，如选题、形成假说、干预措施、实验对象、实验方法等；统计设计是针对统计数据收集进行的设计，如样本来源、样本量、干预措施的分配、统计设计类型测量指标的选择等。统计设计是统计分析的基础，任何设计上的缺陷，都不可能在统计分析阶段弥补和纠正。

[评析]本题考点：研究设计包含的内容。

研究设计是整个研究中最关键的一环，是整个研究过程中始终遵循的依据。正确、严谨、周密的设计是研究工作顺利进行、研究结果真实可靠的保证。因此，应深刻理解并掌握研究设计的内容及其意义。

（四）是非题

描述不确定现象，通过重复观察，发现生物医学领域的不确定现象背后隐藏的统计规律是医学统计的显著特征。（ ）

答案：正确。

[评析] 本题考点：统计方法的特征。

在生物医学研究领域，由于存在较大的生物变异性，并受诸多因素的影响，使实验或观察结果往往成为不确定现象。在大量的重复试验中，这种不确定现象却呈现出明显的统计规律性。统计方法能够帮助人们分析数据，达到去伪存真、去粗存精，透过偶然现象认识其内在的规律性。这正是统计方法的显著特征。

四、习 题

（一） 名词解释

1．总体与样本　2．随机抽样　3．变异　4．等级资料

5．概率与频率　6．随机误差　7．系统误差　8．随机变量

9．参数　10．统计量

（二） 单项选择题

1． 观察单位为研究中的( )。

A．样本　B. 全部对象

C．影响因素　D. 个体

2． 总体是由（ ）。

A．个体组成　B. 研究对象组成

C．同质个体组成　D. 研究指标组成

3． 抽样的目的是（ ）。

A．研究样本统计量　B. 由样本统计量推断总体参数

C．研究典型案例研究误差　D. 研究总体统计量

4． 参数是指（ ）。

A．参与个体数　B. 总体的统计指标

C．样本的统计指标　D. 样本的总和

5． 关于随机抽样，下列那一项说法是正确的（ ）。

A．抽样时应使得总体中的每一个个体都有同等的机会被抽取

B．研究者在抽样时应精心挑选个体，以使样本更能代表总体

C．随机抽样即随意抽取个体

D．为确保样本具有更好的代表性，样本量应越大越好

（三）是非题

1．研究人员测量了100例患者外周血的红细胞数，所得资料为计数资料。

2．统计分析包括统计描述和统计推断。

3．计量资料、计数资料和等级资料可根据分析需要相互转化。

（四）简答题

某年级甲班、乙班各有男生50人。从两个班各抽取10人测量身高，并求其平均身高。如果甲班的平均身高大于乙班，能否推论甲班所有同学的平均身高大于乙班？为什麼？

五、习题答题要点

（一） 名词解释

1．总体：总体（population）是根据研究目的确定的同质的观察单位的全体，更确切的说，是同质的所有观察单位某种观察值（变量值）的集合。总体可分为有限总体和无限总体。总体中的所有单位都能够标识者为有限总体，反之为无限总体。

样本：从总体中随机抽取部分观察单位，其测量结果的集合称为样本（sample）。样本应具有代表性。所谓有代表性的样本，是指用随机抽样方法获得的样本。

2．随机抽样：随机抽样（random sampling）是指按照随机化的原则（总体中每一个观察单位都有同等的机会被选入到样本中），从总体中抽取部分观察单位的过程。随机抽样是样本具有代表性的保证。

3．变异：在自然状态下，个体间测量结果的差异称为变异（variation）。变异是生物医学研究领域普遍存在的现象。严格的说，在自然状态下，任何两个患者或研究群体间都存在差异，其表现为各种生理测量值的参差不齐。

4．等级资料：将观察单位按测量结果的某种属性的不同程度分组，所得各组的观察单位数，称为等级资料(ordinal data)。等级资料又称有序资料。如患者的治疗结果可分为治愈、好转、有效、无效、死亡，各种结果既是分类结果，又有顺序和等级差别，但这种差别却不能准确测量。

5．概率：概率(probability)又称几率，是度量某一随机事件A发生可能性大小的一个数值，记为$P(A)$，$P(A)$越大，说明A事件发生的可能性越大。$0<P(A)<1$。

频率：在相同的条件下，独立重复做n次试验，事件A出现了m次，则比值m/n称为随机事件A在n次试验中出现的频率(freqency)。当试验重复很多次时$P(A)=m/n$。

6．随机误差：随机误差（random error）又称偶然误差，是指排除了系统误差后尚存的误差。它受多种因素的影响，使观察值不按方向性和系统性而随机的变化。误差变量一般服从正态分布。随机误差可以通过统计处理来估计。

7．系统误差：系统误差(systematic error)是指由于仪器未校正、测量者感官的某种偏差、医生掌握疗效标准偏高或偏低等原因，使观察值不是分散在真值的两侧，而是有方向性、系统性或周期性地偏离真值。系统误差可以通过实验设计和完善技术措施来消除或使之减少。

8．随机变量：随机变量（random variable）是指取指不能事先确定的观察结果。随机变量的具体内容虽然是各式各样的，但共同的特点是不能用一个常数来表示，而且，理论上讲，每个变量的取值服从特定的概率分布。

9．参数：参数（paramater）是指总体的统计指标，如总体均数、总体率等。总体参数是固定的常数。多数情况下，总体参数是不易知道的，但可通过随机抽样抽取有代表性的样本，用算得的样本统计量估计未知的总体参数。

10．统计量：统计量（statistic）是指样本的统计指标，如样本均数、样本率等。样本统计量可用来估计总体参数。总体参数是固定的常数，统计量是在总体参数附近波动的随机变量。

（二） 单项选择题

1.D 2.C 3.B 4.B 5.A

（三） 是非题

1．错。外周血的红细胞数是对血液中红细胞含量的测量值，其测量单位为（$10^9/L$），属计量资料。

2．正确。

3．正确。

（四） 简答题

答案：不能。因为，从甲、乙两班分别抽取的10人，测量其身高，得到的分别是甲、乙两班的一个样本。样本的平均身高只是甲、乙两班所有同学平均身高的一个点估计值。既使是按随机化原则进行抽样，由于存在抽样误差，样本均数与总体均数一般很难恰好相等。因此，不能仅凭两个样本均数高低就作出两总体均数熟高熟低的判断，而应通过统计分析，进行统计推断，才能作出判断。

（倪宗瓒 王霞）

第二章　计量资料的统计描述

一、教学大纲要求

（一）掌握内容

1．频数分布表与频数分布图

（1）频数表的编制。

（2）频数分布的类型。

（3）频数分布表的用途。

2．描述数据分布集中趋势的指标

掌握其意义、用途及计算方法。算术均数、几何均数、中位数。

3．描述数据分布离散程度的指标

掌握其意义、用途及计算方法。极差、四分位数间距、方差、标准差、变异系数。

（二）熟悉内容

连续型变量的频数分布图：等距分组、不等距分组。

二、 教学内容精要

计量资料又称为测量资料，它是测量每个观察单位某项指标值的大小所得的资料，一般均有计量单位。常用描述定量资料分布规律的统计方法有两种：一类是用统计图表，主要是频数分布表（图）；另一类是选用适当的统计指标。

（一）频数分布表的编制

频数表（frequency table）用来表示一批数据各观察值或在不同取值区间的出现的频繁程度（频数）。对于离散数据，每一个观察值即对应一个频数，如某医院某年度一日内死亡0，1，2，…20 个病人的天数。如描述某学校学生性别分布情况，男、女生的人数即为各自的频数。对于散布区间很大的离散数据和连续型数据，数据散布区间由若干组段组成，每个组段对应一个频数。制作连续型数据频数表一般步骤如下：

1.求数据的极差（range)。

$$R = X_{\max} - X_{\min} \qquad (2\text{-}1)$$

2.根据极差选定适当“组段”数（通常 8–10 个)。

确定组段和组距。每个组段都有下限 L 和上限 U，数据 X 归组统一定为 $L \leq X < U$ 。

3.写出组段，逐一划记。

频数表可用于揭示资料的分布特征和分布类型，在文献中常用于陈述资料，它便于发现某些特大或特小的可疑值，也便于进一步计算指标和统计分析处理。

（二）描述频数分布中心位置的平均指标

描述中心位置的平均指标，但常因资料的不同而选取不同的指标进行描述。

1.算术均数

算术均数（arithmetic mean）简称均数，描述一组数据在数量上的平均水平。总体均数用 μ 表示，样本均数用 $\bar{X}$ 表示，其计算方法如下：

（1）直接法：直接用原始观测值计算。

$$\bar{X} = \frac{\sum X}{n} \tag{2-2}$$

（2）加权法：在频数表基础上计算，其中 X 为组中值， f 为频数。

$$\bar{X} = \frac{\sum fX}{\sum f} \tag{2-3}$$

2.几何均数

几何均数（geometric mean）用以描述对数正态分布或数据呈倍数变化资料的水平。记为 G。其计算公式为：

(1)直接法

$$G = \lg^{-1}\left(\frac{\sum \lg X}{n}\right) \tag{2-4}$$

（2）加权法

$$G = \lg^{-1}\left(\frac{\sum f \lg X}{\sum f}\right) \tag{2-5}$$

3.中位数

中位数（median）将一组观察值由小到大排列，n 为奇数时取位次居中的变量值；为偶数时，取位次居中的两个变量的平均值。

为奇数时 $M = X_{\left(\frac{n+1}{2}\right)}$ (2-6)

为偶数时 $M = \frac{1}{2}\left(X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}\right)$ (2-7)

2-1 常用平均数的意义及其应用场合

平均数	意义	应用场合
均数	平均数量水平	应用甚广，最适用于对称分布，特别是正态分布
几何均数	平均增（减）倍数	等比资料；对数正态分布
中位数	位次居中的观察值水平	偏态分布；分布不明；分布末端无确定值

（一）反映数据变异程度大小的变异指标

变异指标的应用亦根据资料的不同而选取不同指标进行描述。常用的变异指标有极差、四分位数间距、方差、标准差和变异系数，尤其是方差和标准差更为常用。

1.极差

极差（range）亦称全距，即最大值与最小值之差，用于资料的粗略分析，其计算简便但稳定性较差。

$$R = X_{\max} - X_{\min} \tag{2-1}$$

2.百分位数与四分位数间距

（1）百分位数（percentile）是将 n 个观察值从小到大依次排列，再把它们的位次依次转化为百分位。百分位数的另一个重要用途是确定医学正常参考值范围。 百分位数用 P_x 表示，0

＜x＜100,如25%位数表示为P_{25}。在频数表上，百分位数的计算公式为：

$$P_x = L_x + \frac{i_x}{f_x}(n \cdot x\% - \sum f_L) \quad (2\text{-}8)$$

（2）四分位数间距（inter-quartile range）是由第3四分位数（$Q_3 = P_{75}$）和第1四分位数（$Q_1 = P_{25}$）相减计算而得，常与中位数一起使用，描述偏态分布资料的分布特征，比极差稳定。其计算公式：

$$QR = Q_3 - Q_1 \quad (2\text{-}9)$$

3.方差

方差（variance）表示一组数据的平均离散情况，其计算公式为：

$$S^2 = \frac{\sum (X - m)^2}{n - 1} \quad (2\text{-}10)$$

4.标准差

标准差（standard deviation）是方差的正平方根，使用的量纲与原量纲相同，适用于近似正态分布的资料，大样本、小样本均可，最为常用，其计算公式为：

$$S = \sqrt{\frac{\sum (X - \overline{X})^2}{n - 1}} = \sqrt{\frac{\sum X^2 - (\sum X)^2 / n}{n - 1}} \quad (2\text{-}11)$$

5.变异系数

变异系数（coefficient of variation）用于观察指标单位不同或均数相差较大时两组资料变异程度的比较。用*CV*表示，计算公式为：

$$CV = \frac{S}{\overline{X}} \times 100\% \quad (2\text{-}12)$$

平均指标和变异指标分别反映资料的不同特征，作为资料的总结性统计量，两类指标要求一起使用。如常用$\overline{X} \pm S$或*M*（*QR*）。

三、典型试题分析

1.名词解释：平均数

答案：平均数（average）是描述数据分布集中趋势的指标，在卫生领域中最常用的平均数指标：算术均数、几何均数和中位数。

[评析]本题考察平均数的概念。平均数是一类统计指标，并不单纯指算术均数。

2.描述一组偏态分布资料的变异度，以（　）指标较好。

A.全距　　B.标准差

C.变异系数　　D.四分位数间距

答案：D

[评析]标准差和变异系数均用于描述正态分布资料的变异度，全距和四分位数间距可用于任何资料，而四分位数间距更为稳定，故选D。

3.用均数和标准差可以全面描述（　）资料的特征。

A.正偏态分布　　B.负偏态分布

C.正态分布和近似正态分布　　D.对称分布

答案：C

[评析]本题考察均数和标准差的应用条件。

4.同一资料的标准差是否一定小于均数？

答案：均数和标准差是两类不同性质的统计指标。标准差用于描述数据的变异程度，变异程度大，则该值大，变异程度小，则该值小。标准差可大于均数，也可小于均数。

5.试述极差、四分位数间距、标准差及变异系数的适用范围。

答案：这三个指标均反映计量资料的离散程度。极差与四分位数间距可用于任何分布，后者较前者稳定，但均不能综合反映各观察值的变异程度；标准差最为常用，要求资料近似服从正态分布；变异系数可用于多组资料间度量衡单位不同或均数相差悬殊时的变异程度比较。

四、习　题

（一）名词解释

1.频数表　2.算术均数　3.几何均数　4.中位数　5.极差

6.百分位数　7.四分位数间距　8.方差　9.标准差　10.变异系数

（二）单项选择题

1.各观察值均加（或减）同一数后（　）。

A.均数不变，标准差改变　B.均数改变，标准差不变

C.两者均不变　D.两者均改变

2.比较身高和体重两组数据变异度大小宜采用（　）。

A.变异系数　B.差

C.极差　D.标准差

3.以下指标中（　）可用来描述计量资料的离散程度。

A.算术均数　B.几何均数

C.中位数　D.标准差

4.偏态分布宜用（　）描述其分布的集中趋势。

A.算术均数　B.标准差

C.中位数　D.四分位数间距

5.各观察值同乘以一个不等于0的常数后，（　）不变。

A．算术均数　B.标准差

C.几何均数　D.中位数

6.（　）分布的资料，均数等于中位数。

A.对称　B.左偏态

C.右偏态　D.偏态

7.对数正态分布是一种（　）分布。

A.正态　B.近似正态

C.左偏态　D.右偏态

8.最小组段无下限或最大组段无上限的频数分布资料，可用（　）描述其集中趋势。

A.均数　B.标准差

C.中位数　D.四分位数间距

9.（　）小，表示用该样本均数估计总体均数的可靠性大。

A．变异系数　　B.标准差

C．标准误　　D.极差

10.血清学滴度资料最常用来表示其平均水平的指标是（　）。

A．算术平均数　　B.中位数

C.几何均数　　D．平均数

11.变异系数 *CV*的数值（　）。

A．一定大于 1　　B.一定小于 1

C．可大于 1，也可小于 1　　D.一定比标准差小

12.数列 8、-3、5、0、1、4、-1 的中位数是（　）。

A．2　　B．0

C．2.5　　D．0.5

13.关于标准差,那项是错误的（　）。

A.反映全部观察值的离散程度　　B.度量了一组数据偏离平均数的大小

C.反映了均数代表性的好坏　　D.不会小于算术均数

14.中位数描述集中位置时,下面那项是错误的（　）。

A．适合于偏态分布资料　　B.适合于分布不明的资料

C.不适合等比资料　　D.分布末端无确定值时,只能用中位数

15．5 人的血清滴度为 <1:20、1:40、1:80、1:160、1:320 描述平均滴度，用那种指标较好（　）。

A．平均数　　B.几何均数

C.算术均数　　D．中位数

16.数列 0、48、49、50、52、100 的标准差为（　）。

A．50　　B．26.75

C．28.90　　D．70.78

17.一组变量的标准差将（　）。

A.随变量值的个数 *n* 的增大而增大

B.随变量值的个数 *n* 的增加而减小

C.随变量值之间的变异增大而增大

D.随系统误差的减小而减小

18.频数表计算中位数要求（　）。

A.组距相等　　B.原始数据分布对称

C.原始数据为正态分布或近似正态分布　　D.没有条件限制

19.一组数据中 20%为 3,60%为 2,10%为 1,10%为 0,则平均数为（　）。

A．1.5　　B．1.9

C．2.1　　D．不知道数据的总个数,不能计算平均数

20.某病患者 8 人的潜伏期如下:2、3、3、3、4、5、6、30 则平均潜伏期为（　）。

A.均数为 7 天,很好的代表了大多数的潜伏期

B.中位数为 3 天

C.中位数为 4 天

D.中位数为 3.5 天,不受个别人潜伏期长的影响

21.某地调查 20 岁男大学生 100 名,身高标准差为 4.09cm,体重标准差为 4.10kg,比较两者的变异程度,结果（　）。

A. 体重变异度大

B.身高变异度较大

C.两者变异度相同

D.由单位不同,两者标准差不能直接比较

（三）判断正误并简述理由

1.均数总是大于中位数。（　）

2.均数总是比标准差大。（　）

3.变异系数的量纲和原量纲相同。（　）

4.样本均数大时，标准差也一定会大。（　）

5.样本量增大时，极差会增大。（　）

（四）计算题

1.某卫生防疫站测得大气中的二氧化硫的浓度，用两种计量单位表示：

mg/m^3	：	1	2	3	4	5
ug/m^3	：	1000	2000	3000	4000	5000

分别计算几何均数及标准差，会发现两种不同单位得标准差相等，试解释其原因。

2.尸检中测得北方成年女子 80 人的肾上腺重量（g）如下，试（1）编制频数表，（2）求中位数、均数和标准差。

19.0	12.0	14.0	14.0	8.2	13.0	6.5	12.0	15.0	17.2
12.0	12.7	25.0	8.5	20.0	17.0	8.4	8.0	13.0	15.0
20.0	13.0	13.0	14.0	15.0	7.9	10.5	9.5	10.0	12.0
6.5	11.0	12.5	7.5	14.5	17.5	12.0	10.0	11.0	11.5
16.0	13.0	10.5	11.0	14.0	7.5	14.0	11.4	9.0	11.1
10.0	10.5	8.0	12.0	11.5	19.0	10.0	9.0	19.0	10.0
22.0	9.0	12.0	8.0	14.0	10.0	11.5	11.0	15.0	16.0
8.0	15.0	9.9	8.5	12.5	9.6	18.5	11.0	12.0	12.0

3.测得某地 300 名正常人尿汞值，其频数表如下。试计算均数、中位数、何者代表性较好。

表 2-2 300 例正常人尿汞值（μg/L）频数表

尿汞值	例数	尿汞值	例数	尿汞值	例数
0-	49	24-	16	48-	3
4-	27	28-	9	52-	-
8-	58	32-	9	56-	2
12-	50	36-	4	60-	-
16-	45	40-	5	64-	-

20-	22	44-	-	68-	1

4.有5个变量值7，9，10，14，15，试计算$\overline{X}$及$\sum(X-\overline{X})$。

5.下表为 10 例垂体催乳素微腺瘤经蝶手术前后的血催乳素浓度，试分别求术前、术后的均数，标准差及变异系数。应以何指标比较手术前后数据的变异情况？能说明手术前数据的变异大吗？为什么？

表2-3 手术前后患者血催乳素浓度（mg/ml）

例号	血催乳素浓度		例号	血催乳素浓度	
	术前	术后		术前	术后
1	276	41	6	266	43
2	880	110	7	500	25
3	1600	280	8	1700	300
4	324	61	9	500	215
5	398	105	10	220	92

6.某地微丝蚴血症者 42 例治疗后 7 年用间接荧光抗体试验测得抗体滴度如下。求平均滴度。

抗体滴度的倒数	10	20	40	80	160
例　　数	5	12	13	7	5

五、习题答案要点

（一）名词解释

1.答案：频数表（frequency table）用来表示一批数据各观察值或在不同取值区间的出现的频繁程度（频数）。对于离散数据，每一个观察值即对应一个频数，如某医院某年度一日内死亡0，1，2…20个病人的天数。对于散布区间很大的离散数据和连续型数据，数据散布区间由若干组段组成，每个组段对应一个频数。

2.答案：算术均数（arithmetic mean）描述一组数据在数量上的平均水平。总体均数用μ表示，样本均数用$\overline{X}$表示。

3.答案：几何均数（geometric mean）用以描述对数正态分布或数据呈倍数变化资料的水平。记为G。

4.答案：中位数（median）将一组观察值由小到大排列，*n* 为奇数时取位次居中的变量值；为偶数时，取位次居中的两个变量的平均值。

5.答案：极差（range）亦称全距，即最大值与最小值之差，用于资料的粗略分析，其计算简便但稳定性较差。

6.答案：百分位数（percentile）是将 *n* 个观察值从小到大依次排列，再把它们的位次依次转化为百分位。百分位数的另一个重要用途是确定医学参考值范围。

7.答案：四分位数间距（inter-quartile range）是由第 3 四分位数和第 1 四分位数相减计算而得，常与中位数一起使用，描述偏态分布资料的分布特征，较极差稳定。

8.答案：方差（variance）：方差表示一组数据的平均离散情况，由离均差的平方和除以样本个数得到。

9.答案：标准差（standard deviation）是方差的正平方根，使用的量纲与原量纲相同，适用于近似正态分布的资料，大样本、小样本均可，最为常用。

10.答案：变异系数（coefficient of variation）用于观察指标单位不同或均数相差较大时两组资料变异程度的比较。用 *CV* 表示。

（二）单项选择题

1. B　2.A　3.D　4.C　5.B　6.A　7.C　8.C　9.C　10.C　11.C

12.B　13.D　14.C　15.B　16.C　17.C　18.D　19.B　20.D　21.D

（三）判断正误并简述理由

1.错。均数和中位数的大小关系取决于所描述资料的分布状况。对于负偏态的资料来说，均数大于中位数；对于正偏态的资料来说，均数小于中位数；对称分布的均数和中位数相等。

2.错。

3.错。变异系数无量纲，是一个相对数。

4.错。

5.正确。样本例数越多 ，抽到较大或较小变量值的可能性越大，因而极差可能越大。

（四）计算题

1.答案：用第一组资料计算得几何均数为 2.61 mg/㎥，标准差为 0.27 mg/㎥；第二组资料算得几何均数为 2605.17 ug/㎥，标准差为 0.27 ug/㎥。两组资料均数不等，标准差相等，可见标准差的大小只与资料的离散程度有关，而与均数的大小无关。

2.答案：

（1）编制频数表

①求极差：$R = X_{\max} - X_{\min}$ =25.0-6.5=18.5。

②根据极差确定组距为 2.0，组段数为 10。

③编制频数表。

表 2-4　80 名北方成年女子肾上腺重量（g）频数分布表

肾上腺重量 (g)	组中值 (X)	频　数 (f)	fX	fX^2	累计频数	累计频率 (%)
6.00-	7.00	5	35.00	245.00	5	6.25
8.00-	9.00	14	126.00	1134.00	19	23.75
10.00-	11.00	19	209.00	2299.00	38	47.50
12.00-	13.00	17	221.00	2873.00	55	68.75
14.00-	15.00	12	180.00	2700.00	67	83.75
16.00-	17.00	5	85.00	1445.00	72	90.00
18.00-	19.00	4	76.00	1444.00	76	95.00
20.00-	21.00	2	42.00	882.00	78	97.50
22.00-	23.00	1	23.00	529.00	79	98.75

24.00-	25.00	1	25.00	625.00	80	100.00
合　　计		80	1022.00	14176.00	80	100.00

（2）求中位数，均数和标准差。

①求中位数

$$M = L_x + \frac{i_x}{f_M}\left(\frac{n}{2} - \sum f_L\right) = 12.0 + \frac{2.0}{17}(80 \cdot 50\% - 38) = 12.24g$$

②求均数

$$\overline{X} = \frac{\sum fX}{\sum f} = 12.78$$

③求标准差

$$S = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}} = \sqrt{\frac{\sum fX^2 - ((\sum fX)/\sum f)^2}{\sum f - 1}} = 3.77g$$

3.答案：

表 2-5 300 例正常人尿汞值（μg/L）频数表

尿汞值（μg/L）	组中值（X）	频　数（f）	累计频数	累计频率（%）
0.00-	2.00	49	49	16.33
4.00-	6.00	27	76	25.33
8.00-	10.00	58	134	44.67
12.00-	14.00	50	184	61.33
16.00-	18.00	45	229	76.33
20.00-	22.00	22	251	83.67
24.00-	26.00	16	267	89.00
28.00-	30.00	9	276	92.00
32.00-	34.00	9	285	95.00
36.00-	38.00	4	289	96.33
40.00-	42.00	5	294	98.00
44.00-	46.00	-	294	98.00
48.00-	50.00	3	297	99.00
52.00-	54.00	-	297	99.00
56.00-	58.00	2	299	99.67
60.00-	62.00	-	299	99.67
64.00-	66.00	-	299	99.67
68.00-	70.00	1	300	100.00
合　计		300	300	100.00

（1）求均数

$$\bar{X}=\frac{\sum fX}{\sum f}$$=15.08μg/L

（2）求中位数

$$M=L_x+\frac{i_x}{f_M}\left(\frac{n}{2}-\sum f_L\right)$$=13.28μg/L

由频数表可以看出，此资料为偏态分布，因此用中位数代表性较好。

4.答案：

（1）求均数

$$\bar{X}=\frac{\sum X}{n}=\frac{7+9+10+14+15}{5}=11.00$$

（2）求离均差之和

$$\sum\left(X-\bar{X}\right)$$=0.00

5.答案：

（1）求术前各指标

$$\bar{X}=\frac{\sum X}{n}$$=666.40mg/ml

$$S=\sqrt{\frac{\sum\left(X-\bar{X}\right)^2}{n-1}}$$=551.99mg/ml

$$CV=\frac{S}{\bar{X}}\times100\%$$=82.83%

（2）求术后各指标

$$\bar{X}=\frac{\sum X}{n}$$=127.20 mg/ml

$$S=\sqrt{\frac{\sum\left(X-\bar{X}\right)^2}{n-1}}$$=101.27 mg/ml

$$CV=\frac{S}{\bar{X}}\times100\%$$=79.61% mg/ml

两组资料均数相差悬殊，故而只能用变异系数比较两组何者变异度大，虽然术前变异系数较大，但差异并不明显，需做进一步的统计分析才能知道何者变异度大。

6.答案：

其平均滴度的倒数为

$$G=\lg^{-1}\left(\frac{\sum f\lg X}{\sum f}\right)=\lg^{-1}\left(\frac{65.70}{42}\right)=36.67$$

平均滴度为1∶37。

（姚晨　陈平）

第三章 正态分布

一、教学大纲要求

(一)掌握内容

1．正态分布的概念和特征

（1）正态分布的概念和两个参数；

（2）正态曲线下面积分布规律。

2．标准正态分布

标准正态分布的概念和标准化变换。

3．正态分布的应用

（1）估计频数分布；

（2）制定参考值范围。

(二)熟悉内容

标准正态分布表。

(三)了解内容

1．利用正态分布进行质量控制

2．正态分布是许多统计方法的基础

二、教学内容精要

(一)正态分布

1.正态分布

若 X 的密度函数（频率曲线）为正态函数（曲线）

$$f(X)=\frac{1}{s\sqrt{2p}}e^{-(X-m)^2/(2s^2)} \quad -\infty < X < +\infty \qquad (3\text{-}1)$$

则称 X 服从正态分布，记号 $X\sim N(m,s^2)$。其中 m 、s 是两个不确定常数，是正态分布的参数，不同的 m 、不同的 s 对应不同的正态分布。

正态曲线呈钟型，两头低，中间高，左右对称，曲线与横轴间的面积总等于 1。

2．正态分布的特征

服从正态分布的变量的频数分布由 m 、s 完全决定。

(1) m 是正态分布的位置参数，描述正态分布的集中趋势位置。正态分布以 $x=m$ 为对称轴，左右完全对称。正态分布的均数、中位数、众数相同，均等于 m 。

(2) s 描述正态分布资料数据分布的离散程度，s 越大，数据分布越分散，s 越小，数据分布越集中。s 也称为是正态分布的形状参数，s 越大，曲线越扁平，反之，s 越小，曲线越瘦高。

(二)标准正态分布

1.标准正态分布是一种特殊的正态分布，标准正态分布的 $m=0$，$s^2=1$ ，通常用 u（或 Z）表示服从标准正态分布的变量，记为 $u\sim N$（0，1^2）。

2．标准化变换： $u=\dfrac{X-m}{s}$**，此变换有特性：若** X **服从正态分布** $N(m,s^2)$**，则** u **就服从标准正态分布，故该变换被称为标准化变换。**

3．标准正态分布表

标准正态分布表中列出了标准正态曲线下从-∞到 u **范围内的面积比例** $\Phi(u)$**。**

（三）正态曲线下面积分布

1．实际工作中，正态曲线下横轴上一定区间的面积反映该区间的例数占总例数的百分比，或变量值落在该区间的概率（概率分布）。不同 (X_1,X_2) **范围内正态曲线下的面积可用公式 3-2 计算。**

$$D=\int_{X_1}^{X_2}\frac{1}{s\sqrt{2p}}e^{-(X-m)^2/(2s^2)}dx=\Phi(u_2)-\Phi(u_1) \qquad (3\text{-}2)$$

其中， $u_1=\dfrac{X_1-m}{s}$ **，** $u_2=\dfrac{X_2-m}{s}$**。**

2.几个重要的面积比例

X **轴与正态曲线之间的面积恒等于 1。正态曲线下，横轴区间** $m\pm s$ **内的面积为 68.27%，横轴区间** $m\pm1.64s$ **内的面积为 90.00%，横轴区间** $m\pm1.96s$ **内的面积为 95.00%，横轴区间** $m\pm2.58s$ **内的面积为 99.00%。**

（四）正态分布的应用

某些医学现象，如同质群体的身高、红细胞数、血红蛋白量，以及实验中的随机误差，呈现为正态或近似正态分布；有些指标（变量）虽服从偏态分布，但经数据转换后的新变量可服从正态或近似正态分布，可按正态分布规律处理。其中经对数转换后服从正态分布的指标，被称为服从对数正态分布。

1. **估计频数分布** 一个服从正态分布的变量只要知道其均数与标准差就可根据公式（3-2）估计任意取值 (X_1,X_2) **范围内频数比例。**

2．制定参考值范围

（1）正态分布法 适用于服从正态（或近似正态）分布指标以及可以通过转换后服从正态分布的指标。

（2）百分位数法 常用于偏态分布的指标。表 3-1 中两种方法的单双侧界值都应熟练掌握。

表 3-1 常用参考值范围的制定

概率(%)	正态分布法			百分位数法		
	双侧	单侧		双侧	单侧	
		下限	上限		下限	上限
90	$\overline{X}\pm1.64S$	$\overline{X}-1.28S$	$\overline{X}+1.28S$	$P_5\sim P_{95}$	P_{10}	P_{90}
95	$\overline{X}\pm1.96S$	$\overline{X}-1.64S$	$\overline{X}+1.64S$	$P_{2.5}\sim P_{97.5}$	P_5	P_{95}
99	$\overline{X}\pm2.58S$	$\overline{X}-2.33S$	$\overline{X}+2.33S$	$P_{0.5}\sim P_{99.5}$	P_1	P_{99}

3．质量控制：为了控制实验中的测量（或实验）误差，常以 $\overline{X}\pm2S$ **作为上、下警戒值，以** $\overline{X}\pm3S$ **作为上、下控制值。这样做的依据是：正常情况下测量（或实验）误差服从正态分布。**

4．正态分布是许多统计方法的理论基础。 t **检验、方差分析、相关和回归分析等多种统计方法均要求分析的指标服从正态分布。许多统计方法虽然不要求分析指标服从正态分布，但相应的统计量在大样本时近似正态分布，因而大样本时这些统计推断方法也是以正态分布**

为理论基础的。

三、典型试题分析

1.正态曲线下、横轴上，从均数到$+\infty$的面积为(　　)。

A．95%　　B．50%　　C．97.5%　　D．不能确定（与标准差的大小有关）

答案：B

[评析]　本题考点：正态分布的对称性

因为无论m，s 取什么值，正态曲线与横轴间的面积总等于1，又正态曲线以$X=m$ 为对称轴呈对称分布，所以m 左右两侧面积相等，各为50%。

2．若X服从以m，s 为均数和标准差的正态分布，则X的第95百分位数等于(　　)。

A．$m-1.64s$　　B．$m+1.64s$　　C．$m+1.96s$　　D．$m+2.58s$

答案：B

[评析] 本题考点：正态分布的对称性和面积分布规律

正态分布曲线下$m\pm1.64s$ 范围内面积占90%，则$m\pm1.64s$ 外的面积为10%，又据正态分布的对称性得，曲线下横轴上小于等于$m+1.64s$ 范围的面积为95%，故X的第95百分位数等于$m+1.64s$ 。

3.若正常成人的血铅含量X 近似服从对数正态分布，拟用300名正常人血铅值确定99%参考值范围，最好采用公式(　　)计算。(其中Y=logX)

A. $\overline{X}\pm2.58S$　　B．$\overline{X}+2.33S$

C．$\log^{-1}(\overline{Y}\pm2.58S_Y)$　　D．$\log^{-1}(\overline{Y}+2.33S_Y)$

答案：D

[评析] 本题考点：对数正态分布资料应用正态分布法制定参考值范围

根据题意，正常成人的血铅含量X 近似对数正态分布，则变量X 经对数转换后所得新变量Y应近似服从正态分布，因此可以应用正态分布法估计Y 的99%参考值范围，再求反对数即得正常成人血铅含量X 的99%参考值范围。因血铅含量仅过大为异常，故相应的参考值范围应是只有上限的单侧范围。正态分布法99%范围单侧上限值是均数+2.33倍标准差。

4．正常成年男子红细胞计数近似正态分布,95%参考值范围为3.60～5.84$(\times10^{12}/L)$。若一名成年男子测得红细胞计数为3.10$(\times10^{12}/L)$，则医生判断该男子一定有病。

[评析] 本题考点：参考值范围的涵义

该成年男子不一定有病。因为参考值范围是指绝大多数正常人的指标值范围,故不在此范围内的对象也可能是正常人。

5.假定正常成年女性红细胞数$(\times10^{12}/L)$近似服从均值为4.18，标准差为0.29的正态分布。令X代表随机抽取的一名正常成年女性的红细胞数，求：

(1)　变量X落在区间（4.00，4.50）内的概率；

(2)　正常成年女性的红细胞数95%参考值范围。

[评析] 本题考点：正态分布的应用

(1)根据题意，变量X近似服从正态分布，求变量X落在区间（4.00，4.50）内的概率，即是求此区间内正态曲线下的面积问题，因此，可以把变量X进行标准化变换后，借助标准正态分布表求其面积，具体做法如下：

$$P(4.00<X<4.50)=P(\frac{4.00-4.18}{0.29}<\frac{X-m}{s}<\frac{4.50-4.18}{0.29})$$
$$=P(-0.62<u<1.10)$$
$$=1-\Phi(-1.10)-\Phi(-0.62)$$
$$=1-0.1357-0.2676$$
$$=0.5967$$

变量 X 落在区间(4.00,4.50)内的概率为 0.5967。

(2)问题属于求某个指标的参考值范围问题，因为正常成年女性红细胞数近似服从正态分布，可以直接用正态分布法求参考值范围，又因该指标过高、过低都不正常，所以应求双侧参考值范围，具体做法如下：

下限为：$\overline{X}-1.96s=4.18-1.96(0.29)=3.61(\times10^{12}/L)$

上限为：$\overline{X}+1.96s=4.18+1.96(0.29)=4.75(\times10^{12}/L)$

95%的正常成年女性红细胞数所在的范围是 $3.61\sim4.75(\times10^{12}/L)$。

6.调查得成都市 1979 年 996 名女学生月经初潮年龄的分布如下，本资料宜用何法确定其双侧 99%参考值范围？试估计之。

年岁	10～	11～	12～	13～	14～	15～	16～	17～	18～	19～	20～	合计
人数	7	44	153	244	269	191	61	16	8	1	2	996
累计频率%	0.7	5.1	20.5	45.0	72.0	91.2	97.3	98.9	99.7	99.8	100.0	

[评析] 本题考点：参考值范围的制定

解：本题所给资料明显属于偏态分布资料，所以宜用百分位数法估计其参考值范围。又因此指标过大、过小均属异常，故此参考值范围应是双侧范围。

（1）求 $P_{0.5}$ 首先要找到第 0.5 百分位数所在组，根据累计频率第 0.5 百分位数在第 1 组，因此得 $\sum f_L=0$，$L_X=10$，$f_X=7$，$i_X=1$

代入第二章百分位数的计算公式得：$P_{0.5}=10+\frac{1}{7}(4.98-0)=10.71$（岁）

（2）求 $P_{95.5}$ 先求第 95.5 百分位数所在组为“18～”组，因此得

$$\sum f_L=985，L_X=18，f_X=8，i_X=1$$

代入计算公式得：$P_{95.5}=18+\frac{1}{8}(991.02-985)=18.25$（岁）

成都市女学生月经初潮年龄的双侧 99%参考值范围是 10.71~18.25（岁）。

四、习 题

（一）单项选择题

1.标准正态分布的均数与标准差分别为(　　)。

A．0与1　B．1与0　C．0与0　D．1与1

2.正态分布有两个参数 m 与 s，(　　)相应的正态曲线的形状越扁平。

A．m 越大　B．m 越小　C．s 越大　D．s 越小

3.对数正态分布是一种(　　)分布。

A．正态　B．近似正态　C．左偏态　D．右偏态

4.正态曲线下、横轴上，从均数-1.96 倍标准差到均数的面积为(　　)。

A．95%　　B．45%　　C．97.5%　　D．47.5%

5.标准正态分布曲线下中间90%的面积所对应的横轴尺度u的范围是(　　)。

A．-1.64到+1.64　　B．$-\infty$到+1.64

C．$-\infty$到+1.28　　D．-1.28到+1.28

（二）名词解释

1．正态曲线

2．正态分布

3．标准正态分布

4．标准化变换

（三）简答题

1.简述医学中参考值范围的涵义及制定参考值范围的一般步骤。

2.正态分布、标准正态分布与对数正态分布的联系与区别。

3.对称分布在“$\bar{X} \pm 1.96\bar{S}$标准差”的范围内，也包括95%的观察值吗？

（四）计算题

1.假定 5 岁男童的体重服从正态分布，平均体重m=19.5（kg），标准差s=2.3（kg）。

(1)随机抽查一 5 岁男童的体重，计算概率：

①其体重小于 16.1 kg

②其体重大于 22.9 kg

③其体重在 14.6 kg 到 23.9 kg 之间

(2)试找出最重的5%、10%、2.5% 5 岁男童的体重范围。

2.某年某地测得 200 名正常成人的血铅含量（$mg/100g$）如下，试确定该地正常成人血铅含量的95%参考值范围。

3 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5 6 6 6 6
6 6 6 7 7 7 7 7 7 7 7 7 7 7 7 7 8 8 8 8
8 8 8 8 8 8 8 9 9 9 9 9 9 9 10 10 10 10 10 10
10 10 10 11 11 11 11 11 12 12 12 12 12 12 12 13 13 13 13 13
13 13 13 13 13 13 14 14 14 14 14 14 14 14 14 14 14 15 15 15
15 15 15 16 16 16 16 16 16 17 17 17 17 17 17 17 17 17 17 17
17 17 18 18 18 18 18 19 19 19 19 19 19 20 20 20 20 20 20 20
20 21 21 21 21 21 22 22 22 22 22 22 23 23 23 24 24 24 24 24
24 25 25 26 26 26 26 26 27 27 28 28 29 29 30 30 31 31 31 31
32 32 32 32 32 32 33 33 36 38 38 39 40 41 41 43 47 50 53 60

3.测得某地 300 名正常人尿汞值，其频数表如表 3-2，试用正态分布法和百分位数法估计该地正常人尿汞值的90%，95%，99%上限，讨论用何法估计较适宜。

表 3-2　300 例正常人尿汞值(mg/l)频数表

尿汞值	例数	尿汞值	例数	尿汞值	例数
0～	49	24～	16	48～	3
4～	27	28～	9	52～	-
8～	58	32～	9	56～	2
12～	50	36～	4	60～	-
16～	45	40～	5	64～	-
20～	22	44～	-	68～72	1

4．某市 20 岁男学生 160 人的脉搏数（次/分钟），经正态性检验服从正态分布。求得

$\overline{X}$ = 76.10，S =9.32。试估计脉搏数的95%、99%参考值范围。

5.将测得的238例正常人发汞值(mg/g)从小到大排列，最后14个发汞值如下，求95%单侧上限。

发汞值：	2.6	2.6	2.6	2.6	2.7	2.7	2.7	2.8	2.8	3.0	3.3	4.0	4.1	4.3
秩　次：	225	226	227	228	229	230	231	232	233	234	235	236	237	238

五、习题答题要点

（一）单项选择题

1.A　　2.C　　3. D　　4. D　　5. A

（二）名词解释

1．正态曲线：正态曲线（normal curve）是函数

$$f(X)=\frac{1}{s\sqrt{2p}}e^{-(X-m)^2/(2s^2)}\quad ,\ -\infty<X<+\infty$$

对应的曲线。此曲线呈钟型，两头低中间高，左右对称。

2．正态分布：若指标X的频率曲线对应于数学上的正态曲线，则称该指标服从正态分布(normal distribution)。通常用记号$N(m,s^2)$表示均数为m，标准差为s的正态分布。

3.标准正态分布:均数为0、标准差为1的正态分布被称为标准正态分布(standard normal distribution)，通常记为$N(0,1^2)$。

4．标准化变换：$u=\frac{X-m}{s}$，此变换有特性：若X服从正态分布$N(m,s^2)$，则u就服从标准正态分布，故该变换被称为标准化变换（standardized transformation)。

（二）简答题

1．医学中常把绝大多数正常人的某指标范围称为该指标的参考值范围，也叫正常值范围。所谓“正常人”不是指完全健康的人，而是指排除了所研究指标的疾病和有关因素的同质人群。

制定参考值范围的一般步骤：

(1) 定义“正常人”，不同的指标“正常人”的定义也不同。

(2) 选定足够数量的正常人作为研究对象。

(3) 用统一和准确的方法测定相应的指标。

(4) 根据不同的用途选定适当的百分界限，常用95%。

(5) 根据此指标的实际意义，决定用单侧范围还是双侧范围。

(6) 根据此指标的分布决定计算方法，常用的计算方法：正态分布法、百分位数法。

2．三种分布均为连续型随机变量的分布。正态分布、标准正态分布均为对称分布，对数正态分布是不对称的，其峰值偏在左边。标准正态分布是一种特殊的正态分布（均数为0，标准差为1)。一般正态分布变量经标准化转换后的新变量服从标准正态分布。对数正态分布不属于正态分布的范畴，对数正态分布变量经对数转换后的新变量服从正态分布。

3.不一定。均数$\pm$1.96标准差范围内包含95%的变量值是正态分布的分布规律，不是对称分布的规律。对称分布不一定是正态分布。

（三）计算题：

1.解：(1) 设该男童的体重为X kg,则

① $P(X<16.1)=P(\frac{X-19.5}{2.3}<\frac{16.1-19.5}{2.3})=P(u<-1.48)=\Phi(-1.48)=0.0694$

② $P(X>22.9)=1-P(X\leq 22.9)=1-P(\frac{X-19.5}{2.3}\leq\frac{22.9-19.5}{2.3})=1-P(u\leq 1.48)=\Phi(-1.48)=0.0694$

③ $P(14.6\leq X\leq 23.9)=P(X\leq 23.9)-P(X\leq 14.6)$

$=P(\frac{X-19.5}{2.3}\leq\frac{23.9-19.5}{2.3})-P(\frac{X-19.5}{2.3}\leq\frac{14.6-19.5}{2.3})$

$=P(u\leq 1.91)-P(u\leq -2.13)$

$=1-\Phi(-1.91)-\Phi(-2.13)$

$=0.9719-0.0166=0.9553$

（2）设最重的5%，10%，2.5%男童体重的下限分别为x_1 kg，x_2 kg，x_3 kg

$P(X>x_1)=0.05$　　$P(u\leq\frac{x_1-19.5}{2.3})=0.95$

又∵$P(u\leq 1.645)=0.95$　　$\therefore\frac{x_1-19.5}{2.3}=1.645$　　$x_1=23.3$（kg）

$P(X>x_2)=0.10$　　因为正态分布关于均数对称，所以

$$P(\frac{X-19.5}{2.3}>\frac{x_2-19.5}{2.3})=P(\frac{X-19.5}{2.3}<-\frac{x_2-19.5}{2.3})=P(u<-\frac{x_2-19.5}{2.3})=\Phi(-\frac{x_2-19.5}{2.3})=0.10$$

查标准正态曲线下面积表$-\frac{x_2-19.5}{2.3}=-1.282$　　故 $x_2=22.4$（kg）

同理　$x_3=24.0$（kg）

2．解：正常成人的血铅含量近似对数正态分布，经对数转换后应近似服从正态分布，所以对原始数据作对数变换，并编制频数表，再利用正态分布法求95%参考值范围。对数换算过程如表3-3所示。

表3-3　200名正常成人血铅含量（m g /100g）对数值频数表

对数组段	真数组段	频数
0.45—	3—	1
0.55—	4—	5
0.65—	5—	10
0.75—	6—	20
0.85—	8—	11
0.95—	9—	21
1.05—	12—	29
1.15—	15—	25
1.25—	18—	30
1.35—	23—	20
1.45—	29—	16
1.55—	36—	8
1.65—	45—	3
1.75—1.85	57—	1
		200

依据表3-3，设x为对数组段的组中值，n=200，$\sum fx$=230，$\sum fx^2$=279.04

则$\overline{X}=\frac{\sum fx}{n}=\frac{279.04}{200}=1.15$（m mol/L）

$$S=\sqrt{\frac{\sum fx^2-(\sum fx)^2/n}{n-1}}=\sqrt{\frac{279.04-(230)^2/200}{200-1}}=0.2703\ (m\ mol/L)$$

该地正常成人血铅含量为对数正态分布，按正态分布法估计参考值范围，又因此指标过大属异常，故此参考值范围应为单侧范围。

故单侧95%上限为： $\log^{-1}(\overline{X}+1.64S_X)=\log^{-1}(1.15+1.64\times0.2703)=39\ (m\ mol/L)$

所以该地正常成人血铅含量95%参考值范围上限为39（m mol/L）。

3．解：由表3-2得300名正常人尿汞值$\overline{X}=15.08(mg/l)$，S=11.10(mg/l)

用正态分布法估计正常值范围：

90%正常值范围上限为：$\overline{X}+1.28S$=15.08+1.28（11.10）=29.29(mg/l)

95%正常值范围上限为：$\overline{X}+1.64S$=15.08+1.64（11.10）=33.28(mg/l)

99%正常值范围上限为：$\overline{X}+2.33S$=15.08+2.33（11.10）=40.94(mg/l)

用百分位数法估计正常值范围：

90%正常值范围上限为：$P_{90}=28+\frac{4}{9}(300\times90\%-267)$=29.33$(mg/l)$

90%正常值范围上限为：$P_{95}=36+\frac{4}{9}(300\times95\%-285)$=36.00$(mg/l)$

99%正常值范围上限为：$P_{99}=52+\frac{4}{9}(300\times99\%-297)$=52.00$(mg/l)$

本题正常人尿汞值属于偏态分布资料，用百分位数法估计较适宜。

4．解：脉搏数的95%正常值范围为：$\overline{X}\pm1.96S$=76.10±1.96（9.32）=57.83～94.37

脉搏数的99%正常值范围为：$\overline{X}\pm2.58S$=76.10±2.58（9.32）=52.05～100.37

5．解：（238+1）×0.95=227.05，则95%上限即为第227个数据与第228个数据之间。因为第227个和第228个数据均为2.6，故95%正常值范围的上限应为2.6(mg/g)。

（曹素华　杜晓晗）

第四章　总体均数的估计和假设检验

一、教学大纲要求

（一） 掌握内容

1．抽样误差、可信区间的概念及计算；

2．总体均数估计的方法；

3．两组资料均数比较的方法，理解并记忆应用这些方法的前提条件；

4．假设检验的基本原理、有关概念（如Ⅰ、Ⅱ类错误）及注意事项。

（二） 熟悉内容

两样本方差齐性检验。

（三） 了解内容

1．*t*分布的图形与特征；

2．总体方差不等时的两样本均数的比较；

3．等效检验。

二、教学内容精要

（一） 基本概念

1．抽样误差

抽样研究中，样本统计量与总体参数间的差别称为抽样误差（sampling error）。统计上用标准误（standard error，SE）来衡量抽样误差的大小。不同的统计量，标准误的表示方法不同，如均数的标准误用 $S_{\bar{X}}$ 表示，率的标准误用 S_p 表示，回归系数的标准误用 S_b 表示等等。均数的标准误与标准差的区别见表4-1。

表4-1　均数的标准误与标准差的区别

	均数的标准误	标准差
意义	反映 $\bar{X}$ 的抽样误差大小	反映一组数据的离散情况
记法	$s_{\bar{X}}$（样本估计值 $S_{\bar{X}}$）	s（样本估计值 S）
计算	$s_{\bar{X}}=\frac{s}{\sqrt{n}}$ $S_{\bar{X}}=\frac{S}{\sqrt{n}}$	$s=\sqrt{\frac{\sum(X-m)^2}{n}}$ $S=\sqrt{\frac{\sum(X-\bar{X})^2}{n-1}}$
控制方法	增大样本含量可减小标准误。	个体差异或自然变异，不能通过统计方法来控制。

2．可信区间

（1）定义、涵义：即按预先给定的概率确定的包含未知总体参数的可能范围。该范围称为总体参数的可信区间（confidence interval，*CI*）。它的确切含义是：*CI* 是随机的，总体参数是固定的，所以，*CI* 包含总体参数的可能性是 1-α 。不能理解为 *CI* 是固定随机的，总体参数是随机固定的，总体参数落在 *CI* 范围内可能性为 1-α 。当$\alpha = 0.05$时，称为 95%可信区间，记作95%*CI*。当$\alpha = 0.01$时，称为99%可信区间，记作99%*CI*。

（2）可信区间估计的优劣：一定要同时从可信度（即 1-α 的大小）与区间的宽度两方面来衡量。

(二) *t* 分布与正态分布

t 分布与标准正态分布相比有以下特点：①都是单峰、对称分布；②*t* 分布峰值较低，而尾部较高；③随自由度增大，*t* 分布趋近与标准正态分布；当$n \to \infty$时，*t* 分布的极限分布是标准正态分布。

(三)总体均数的估计

参数估计有点估计和区间估计两种方式。总体均数的估计,见表 4-2。

表 4-2　总体均数的估计

	点估计	区间估计
意义	直接用样本统计量代替总体参数。	用统计量$\overline{X}$和$S_{\overline{x}}$确定一个有概率意义的区间，以该区间具有较大的可信度包含总体均数。
估计方法	以$\overline{X}$作为估计值	①小样本（$\overline{X}-t_{\alpha/2,n}S_{\overline{x}}$，$\overline{X}+t_{\alpha/2,n}S_{\overline{x}}$） ②大样本（$\overline{X}-u_{\alpha/2}S_{\overline{x}}$，$\overline{X}+u_{\alpha/2}S_{\overline{x}}$） ③两总体均数差值的可信区间 （$\overline{X}_1-\overline{X}_2-t_{\alpha/2,n}S_{\overline{x}_1-\overline{x}_2}$，$\overline{X}_1-\overline{X}_2+t_{\alpha/2,n}S_{\overline{x}_1-\overline{x}_2}$）

(四)两均数差别的比较

1．**样本均数和总体均数比较的 *t* 检验**

前提：服从正态分布

H_0：$m = m_0$；H_1：$m \neq m_0$

$$t=\frac{\overline{X}-m_0}{S_{\overline{X}}}，\ n=n-1 \qquad (4\text{-}1)$$

2．配对设计的 *t* 检验

前提：差值服从正态分布

H_0：$m_d = 0$；H_1：$m_d \neq 0$

$$t=\frac{\overline{d}-m_d}{S_{\overline{d}}}，\quad n=n-1 \qquad (4\text{-}2)$$

3．成组设计的两样本均数比较的 *t* 检验

前提：两组数据均服从正态分布；两组总体方差相等

H_0：$m_1 = m_2$；H_1：$m_1 \neq m_2$

$$t=\frac{\overline{X}_1-\overline{X}_2}{S_{\overline{X}_1-\overline{X}_2}},\ n=n_1+n_2-2 \qquad (4\text{-}3)$$

其中，$S_{\overline{X}_1-\overline{X}_2}=\sqrt{S_c^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$，$S_c^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ (4-4)

$S_{\overline{X}_1-\overline{X}_2}$ 表示两样本均数差值的标准误。

4．单样本 *u* 检验

前提：当样本较大（如 n>50）或总体 s_0 已知时

$$u=\frac{\overline{X}-m_0}{S/\sqrt{n}}\ （n\text{ 较大时}） \qquad (4\text{-}5)$$

$$u=\frac{\overline{X}-m_0}{s_0/\sqrt{n}}\ （s_0\text{ 已知时}） \qquad (4\text{-}6)$$

5．大样本均数比较的 *u* 检验

前提：样本足够大

成组设计的两样本均数比较可用：

$$u=\frac{\overline{X}_1-\overline{X}_2}{\sqrt{S_{\overline{x}_1}^2+S_{\overline{x}_2}^2}} \qquad (4\text{-}7)$$

6．要推断组间没有差别或差别很小，应采用等效检验（squivalence test）。

(五)假设检验的步骤及有关概念

1．基本思想：把握“小概率事件在一次抽样试验中是几乎不可能发生”的原理。

2．步骤：①建立假设、选用单侧或双侧检验、确定检验水准；②选用适当检验方法，计算统计量；③确定 *P* 值并作出推断结论。

3．Ⅰ类错误：H_0 为真（实际无差别），假设检验结果拒绝 H_0，接受 H_1（推论有差别）所犯的错误称为Ⅰ类错误（type Ⅰ error），Ⅰ类错误的概率记作 a。

Ⅱ类错误：H_1 为真（实际有差别），假设检验结果拒绝 H_1，接受 H_0（推论无差别）所犯的错误称为Ⅱ类错误（type Ⅱ error），Ⅱ类错误的概率记作 b。

4．1- b 称为检验效能，过去称把握度（power of test），即两总体确有差别，按 a 水准能发现该差别的能力。

三、典型试题分析

（一） 单项选择题

1．当样本含量增大时，以下说法正确的是（　　）

A. 标准差会变小

B. 样均数标准误会变小

C. 均数标准误会变大

D．标准差会变大

答案：B

[评析] 本题考点：这道题是考察均数标准误的概念。

从均数标准误的定义讲，它反映的是均数抽样误差的大小，那么样本含量越大，抽样误差应该越小。从均数标准误的计算公式 $S_{\bar{x}}=S/\sqrt{n}$ 来看，也应是 n 越大，$S_{\bar{x}}$ 越小。

2．区间 $\overline{X}\pm 2.58S_{\overline{x}}$ 的含义是（　　）

A．99%的总体均数在此范围内　　B．样本均数的 99%可信区间

C．99%的样本均数在此范围内　　D．总体均数的 99%可信区间

答案：D

[评析] 本题考点：可信区间的含义。

可信区间的确切含义指的是：总体参数是固定的，可信区间包含了总体参数的可能性是 $1-\alpha$，而不是总体参数落在 CI 范围的可能性为 $1-\alpha$。本题 B、D 均指样本均数，首先排除。A 说总体均数在此范围内，显然与可信区间的含义相悖。因此答案为 D。

（二） 是非题

1．进行两均数差别的假设检验时，当 P=0.05 时，则拒绝 H_0；当 P>0.05 时，则接受 H_0，认为两总体均数无差别。

[评析] 答案：错误。当 P=0.05，拒绝 H_0 时，我们是依据 α 这一小概率来下结论的。而当 P>0.05 时，我们对两总体均数无差别这一结论无任何概率保证，因此不能贸然下无差别的结论。正确的说法是，按所取检验水准 α，接受 H_1 的统计证据不足。

2．通常单侧检验较双侧检验更为灵敏，更易检验出差别，应此宜广泛使用。

[评析] 答案：错误。根据专业知识推断两个总体是否有差别时，是甲高于乙，还是乙高于甲，当两种可能都存在时，一般选双侧；若根据专业知识，如果甲不会低于乙，或者研究者仅关心其中一种可能时，可选用单侧。一般来讲，双侧检验较为稳妥。单侧检验，应以专业知识为依据，它充分利用了另一侧的不可能性，故检出率高，但应慎用。

3．只要增加样本含量到足够大，就可以避免 I 和 II 型错误。

[评析] 答案：错误。因为通过假设检验推断出的结论具有概率性，因此出现错误判断的可能性就一定存在，无论用任何方法也不能消除这一可能。但是，我们可以使错误判断的可能性尽量地小，比如样本含量越大，犯 I 和 II 类错误的可能性越小。

（三） 简答题

1． 简述可信区间在假设检验问题中的作用。

[评析] 可信区间不仅能回答差别有无统计学意义，而且还能提示差别有无实际意义。可信区间只能在预先规定的概率即检验水准 α 的前提下进行计算，而假设检验能够获得一较为确切的概率 P 值。故将二者结合起来，才是对假设检验问题的完整分析。

2．某医生就 4-3 资料，对比用胎盘浸液钩端螺旋体菌苗对 328 名农民接种前、后血清抗体的变化。

表 4-3　328 名农民血清抗体滴度及统计量

	抗体滴度的倒数								$\overline{X}$	S	$S_{\overline{x}}$
	0	20	40	80	160	320	640	1280			
免疫前人数	211	27	19	24	25	19	3	0	76.1	111.7	6.17
免疫后人数	2	16	57	76	75	54	25	23	411.9	470.5	25.90

t =（411.91-76.10）/$\sqrt{25.90^2+6.17^2}$ =12.6，按 $n=14$ 查 t 界值表 P<0.01，说明接种后血清抗体有增长。

问该医生在整理资料和分析资料过程中有何不妥？

答： ①资料整理不当，未整理成配对资料；②统计描述指标使用不当，对于滴度的倒数不宜用算术均数、标准差，有“0”出现，也不宜算几何均数。比较免疫前后抗体滴度的倒数，应计算中位数和四分位数间距；③不宜用 t 检验。可将抗体滴度的倒数经对数或平方根转换后，做配对 t 检验（ν=327）。

（四） 计算题

1．某医院用新药与常规药物治疗婴幼儿贫血，将 20 名贫血患儿随机等分两组，分别接受两种药物治疗，测得血红蛋白增加量（g/L）见表 4-4。问新药与常规药的疗效有无差别？

表 4-4 两种药物治疗婴幼儿贫血结果

治疗药物	血红蛋白增加量（g/L）									
新药组	24	36	25	14	26	34	23	20	15	19
常规药组	14	18	20	15	22	24	21	25	27	23

解：本题属成组设计资料。

$H_0: m_1 = m_2$　　$H_1: m_1 \neq m_2$　　$a = 0.05$

$t = \dfrac{\overline{X}_1 - \overline{X}_2}{S_{\overline{x}_1 - \overline{x}_2}}$，$n = n_1 + n_2 - 2$

$t = \dfrac{2.7}{2.6485} = 1.019$，$n = 18$

P>0.05

因此，根据现有资料尚不能认为新药与常规药的疗效有差别。

2.将 20 名某病患者随机分为两组，分别用甲、乙两药治疗，测得治疗前后的血沉（mm/h）见表 4-5。问：（1）甲、乙两药是否均有效？（2）甲、乙两药疗效是否有别？

表 4-5 甲、乙两药治疗某病情况

	序号	1	2	3	4	5	6	7	8	9	10
甲药	治疗前	30	33	26	31	30	27	28	28	25	29
	治疗后	26	29	23	30	30	24	22	25	23	23
	序号	11	12	13	14	15	16	17	18	19	20
乙药	治疗前	29	30	29	33	28	26	30	31	30	30
	治疗后	26	23	25	23	23	25	28	22	27	24

（1）解：对甲、乙两药治疗数据分别采用配对 t 检验，得

甲药：$t = \overline{d} / S_{\overline{d}} = 3.2/0.611 = 5.237$

乙药：$t = \overline{d} / S_{\overline{d}} = 5.0/0.9428 = 5.303$

ν=9，P<0.001，按a =0.05 水准，拒绝 H_0，接受H_1，故可认为甲乙两药治疗前后均有差别。

（2）解：由表中资料分别求得治疗前后差值，再做两组比较。

$t=\frac{\bar{d}_1-\bar{d}_2}{S_{\bar{d}_1-\bar{d}_2}}$=-1.602，$v$=18，得 0.2>$P$>0.1，按$\alpha$ =0.05 水准，不拒绝 H_0，尚不能认为甲、乙两药疗效有差别。

3．测得某地 90 名正常成年女性红细胞数（10^4/mm^3）的均值 418、标准差为 29。试求：

（1） 该地 95%的正常成年女性红细胞数所在的范围；

（2） 该地正常成年女性红细胞数总体均数的 95%可信区间。

解：(1）用正态分布法估计正常值范围。因红细胞过多或过少均为异常,故此参考值范围应是双侧范围。

上限：$\bar{X}$ +1.96S=418+1.96× 29=474.84（10^4/mm^3）

下限：$\bar{X}$ -1.96S=418-1.96× 29=361.16（10^4/mm^3）。

即（361.16，474.84）（10^4/mm^3）。

（2）由于 n=90>50，故可近似为正态分布。

上限：$\bar{X}$ +1.96$S_{\bar{X}}$ =418+1.96× 29/$\sqrt{90}$ =423.99（10^4/mm^3）

下限：$\bar{X}$ -1.96$S_{\bar{X}}$ =418-1.96× 29/$\sqrt{90}$ =412.01（10^4/mm^3）。

即（412.01，423.99）（10^4/mm^3）。

四、习　题

（一）　单项选择题

6． 标准误的英文缩写为：

A．S　　B．SE　　C．$S_{\bar{X}}$　　D．SD

7． 通常可采用以下那种方法来减小抽样误差：

A．减小样本标准差　　B．减小样本含量

C．扩大样本含量　　D．以上都不对

8． 配对设计的目的：

A．提高测量精度　　B．操作方便

C．为了可以使用 t 检验　　D．提高组间可比性

9． 以下关于参数估计的说法正确的是：

A． 区间估计优于点估计

B． 样本含量越大，参数估计准确的可能性越大

C． 样本含量越大，参数估计越精确

D．对于一个参数只能有一个估计值

10．关于假设检验，下列那一项说法是正确的

A．单侧检验优于双侧检验

B．采用配对 t 检验还是成组 t 检验是由实验设计方法决定的

C．检验结果若 P 值大于 0.05，则接受 H_0 犯错误的可能性很小

D．用 u 检验进行两样本总体均数比较时，要求方差齐性

6. 两样本比较时，分别取以下检验水准，下列何者所取第二类错误最小

A．α =0.05　　B．α =0.01　　C．α =0.10　　D．α =0.20

7. 统计推断的内容是

A．用样本指标推断总体指标　　B．检验统计上的“假设”

C．A、B均不是　　D．A、B均是

8．当两总体方差不齐时，以下哪种方法不适用于两样本总体均数比较

A．t检验　　B．t'检验

C．u 检验（假设是大样本时）　　D．F检验

9．甲、乙两人分别从随机数字表抽得30个（各取两位数字）随机数字作为两个样本，求得$\overline{X}_1$，S_1^2，$\overline{X}_2$，S_2^2，则理论上

A．$\overline{X}_1=\overline{X}_2$，$S_1^2=S_2^2$

B．作两样本t检验，必然得出无差别的结论

C．作两方差齐性的F检验，必然方差齐

D．分别由甲、乙两样本求出的总体均数的95%可信区间，很可能有重叠

10．以下关于参数点估计的说法正确的是

A．CV越小，表示用该样本估计总体均数越可靠

B．$s_{\overline{X}}$越小，表示用该样本估计总体均数越准确

C．$s_{\overline{X}}$越大，表示用该样本估计总体均数的可靠性越差

D．S越小，表示用该样本估计总体均数越可靠

（二） 名词解释

1．统计推断

2．抽样误差

3．标准误及$s_{\overline{X}}$

4．可信区间

5．参数估计

6．假设检验中P的含义

7．I型和II型错误

8．检验效能

9．检验水准

（三） 是非题

1．若两样本均数比较的假设检验结果P值远远小于0.01，则说明差异非常大。

2．对同一参数的估计，99%可信区间比90%可信区间好。

3．均数的标准误越小，则对总体均数的估计越准确。

（四） 简答题

1．假设检验时，当$P\leq0.05$，则拒绝H_0，理论依据是什么？

2．假设检验中α与P的区别何在？

（五） 计算题

1. 治疗10名高血压病人，对每一种病人治疗前、后的舒张压（mmHg）进行了测量，结果见（表4-6），问治疗前后有无差异？

表4-6　10名高血压病人治疗前后的舒张压（mmHg）

病例编号	1	2	3	4	5	6	7	8	9	10

治疗前	117	127	141	107	110	114	115	138	127	122
治疗后	123	108	120	107	100	98	102	152	104	107

2.某医院病理科研究人体两肾的重量，20 例男性尸解时的左、右肾的称重记录见表 4-7，问左、右肾重量有无不同？

表 4-7　20 例男性尸解时左、右肾的称重记录

编号	左肾（克）	右肾（克）
1	170	150
2	155	145
3	140	105
4	115	100
5	235	222
6	125	115
7	130	120
8	145	105
9	105	125
10	145	135
11	155	150
12	110	125
13	140	150
14	145	140
15	120	90
16	130	120
17	105	100
18	95	100
19	100	90
20	105	125

3．有 13 例健康人，11 例克山病人的血磷测定值（mg%）如表 4-8 所示，问克山病人的血磷是否高于健康人？

表 4-8　健康人与克山病人的血磷测定值（mg%）

健康人	170	155	140	115	235	125	130	145	105	145
患　者	150	125	150	140	90	120	100	100	90	125

2．某生化实验室测定了几组人的血清甘油三酯含量（mg%）见表 4-9，试分析比较工人与干部，男与女的该项血酯水平。

表 4-9　正常成人按不同职业、性别分类的的血清甘油三酯含量（mg%）

	人数	平均数	标准差
工人	112	106.49	29.09
干部	106	95.93	26.63
男	116	103.91	27.96
女	102	97.93	28.71

五、习题答题要点

（四）　单项选择题

1.B　2.C　3.D　4.B　5.B　6.D　7.D　8.A　9.D　10.C

（五）　名词解释

1．统计推断：通过样本指标来说明总体特征，这种从样本获取有关总体信息的过程称为统计推断（statistical inference）。

2．抽样误差：由个体变异产生的，抽样造成的样本统计量与总体参数的差异，称为抽样误差（sampling error）。

3．标准误及$s_{\bar{X}}$：通常将样本统计量的标准差称为标准误。许多样本均数的标准差$s_{\bar{X}}$称为均数的标准误（standard error of mean，SEM），它反映了样本均数间的离散程度，也反映了样本均数与总体均数的差异，说明均数抽样误差的大小。

4．可信区间：按预先给定的概率确定的包含未知总体参数的可能范围。该范围称为总体参数的可信区间（confidence interval，CI）。它的确切含义是：可信区间包含总体参数的可能性是 1-a，而不是总体参数落在该范围的可能性为 1-a。

5．参数估计：指用样本指标值（统计量）估计总体指标值（参数）。参数估计有两种方法：点估计和区间估计。

6．假设检验中 P 的含义：指从 H_0 规定的总体随机抽得等于及大于（或等于及小于）现有样本获得的检验统计量值的概率。

7．I 型和 II 型错误：I 型错误（type I error），指拒绝了实际上成立的H_0，这类“弃真”的错误称为 I 型错误，其概率大小用 a 表示；II 型错误（type II error），指接受了实际上不成立的 H_0，这类“存伪”的误称为 II 型错误，其概率大小用 b 表示。

8．检验效能：1-b 称为检验效能（power of test），它是指当两总体确有差别，按规定的检验水准a 所能发现该差异的能力。

9．检验水准：，是预先规定的，当假设检验结果拒绝 H_0，接受 H_1，下“有差别”的结论时犯错误的概率称为检验水准（level of a test），记为a。

（六）　是非题

1．错。P 值的大小只能说明差异是否有统计学意义，同样的差异，例数越多，P 值越小。

2．错。可信区间的优劣要通过两点衡量：区间的可信度；区间的宽度。因此不能笼统的通过区间可信度的大小来评价优劣。

3． 正确。标准误越小，可信区间越窄，对总体均数估计的准确程度越高。

（七） 简答题

1．答：P 值系由 H_0 所规定的总体做随机抽样，获得等于及大于（或等于及小于）依据现有样本信息所计算得的检验统计量的概率。

当 $P \leq 0.05$ 时，说明在 H_0 成立的条件下，得到现有检验结果的概率小于 α ，因为小概率事件几乎不可能在一次试验中发生，所以拒绝 H_0。同时，下“有差别”的结论的同时，我们能够知道可能犯错误的概率不会大于 α ，也就是说，有了概率保证。

2．答：以 t 检验为例，α 与 P 都可用 t 分布尾部面积大小表示，所不同的是：α 值是指在统计推断时预先设定的一个小概率值，就是说如果 H_0 是真的，允许它错误的被拒绝的概率。P 值是由实际样本获得的，是指在 H_0 成立的前提下，出现等于或大于现有检验统计量的概率。

（八） 计算题

1．解：本题属配对设计资料，故应用配对 t 检验方法计算。t=2.484，ν=9，P<0.05，按 α =0.05 水准拒绝 H_0，认为治疗前后有差别（注：此类研究是非随机的自身前后对比研究，要确认疗效，应设立平行对照）。

2．解：本题属配对设计资料，故应用配对 t 检验方法计算。t=2.157，ν=19，P<0.05，按 α =0.05 水准拒绝 H_0，认为左、右肾重量差别有统计学意义，右较左肾轻。

3．解：本题属成组设计资料，故应用成组 t 检验方法计算。t=2.539，ν=22，P<0.05，按 α =0.05 水准拒绝 H_0，认为二者血磷含量差别有统计学意义，克山病人的血磷高于健康人。（注：此类研究是非随机化的对比研究，如果病人与健康人不具可比性，如居住地不同、性别不同、年龄不同，则不能保证结论正确。

4．解：本题可通过计算两均数差值的 95% 或 99% 可信区间来判断两总体均数的差别。

工人与干部均数差值的 95%和 99%可信区间分别为：(3.10，18.02)，(0.73，20.39)，均不包含 0 在内，故可认为工人与干部血清甘油三酯含量的总体均属有差别。

男性与女性均数差值的 95%和 99%可信区间分别为：(-1.60，13.56)，(-4.01，15.97)，均包含 0 在内，故尚不能认为男性与女性血清甘油三酯含量的总体均属有差别。

（潘晓平 马跃渊）

第五章 方差分析

一、教学大纲要求

（一）掌握内容

1．方差分析基本思想

（1） 多组计量资料总变异的分解，组间变异和组内变异的概念。

（2） 多组均数比较的检验假设与 *F* 值的意义。

（3） 方差分析的应用条件。

2．常见实验设计资料的方差分析

（1）完全随机设计的单因素方差分析：适用的资料类型、总变异分解（包括自由度的分解）、方差分析的计算、方差分析表。

（2）随机区组设计资料的两因素方差分析：适用的资料类型、总变异分解（包括自由度的分解）、方差分析的计算、方差分析表。

（3）多个样本均数间的多重比较方法： LSD-t **检验法；**Dunnett-t **检验法；**SNK-q **检验法。**

（二）熟悉内容

多组资料的方差齐性检验、变量变换方法。

（三）了解内容

两因素析因设计方差分析、重复测量设计资料的方差分析。

二、教学内容精要

(一) 方差分析的基本思想

1． 基本思想

方差分析（analysis of variance，ANOVA **）的基本思想就是根据资料的设计类型，即变异的不同来源将全部观察值总的离均差平方和（**sum of squares of deviations from mean，SS**）和自由度分解为两个或多个部分，除随机误差外，其余每个部分的变异可由某个因素的作用（或某几个因素的交互作用）加以解释，如各组均数的变异 $SS_{组间}$ 可由处理因素的作用加以解释。通过各变异来源的均方与误差均方比值的大小，借助 F 分布作出统计推断，判断各因素对各组均数有无影响。**

2．分析三种变异

（1）组间变异：各处理组均数之间不尽相同，这种变异叫做组间变异（variation among groups**），组间变异反映了处理因素的作用（处理确有作用时 ），也包括了随机误差（ 包括个体差异及测定误差 ），其大小可用组间均方（$MS_{组间}$）表示，即 $MS_{组间}= SS_{组间}/n_{组间}$ ， 其中，$SS_{组间}=\sum_{i=1}^{k} n_i(\bar{x}_i-\bar{x})^2$ ，$n_{组间}=k\text{-}1$ 为组间自由度。k 表示处理组数。**

（2）组内变异：各处理组内部观察值之间不尽相同，这种变异叫做组内变异 (variation within groups**)，组内变异反映了随机误差的作用，其大小可用组内均方 （$MS_{组内}$）表示，$MS_{组内}=SS_{组内}/n_{组内}$ ，其中 $SS_{组内}=\sum_{i=1}^{k}\left[\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2\right]$，$n_{组内}=N-k$ ，为组内均方自由度。**

（3）总变异：所有观察值之间的变异（不分组），这种变异叫做总变异(total variation**)。**

其大小可用全体数据的方差表示， 也称总均方($MS_{总}$)。按方差的计算方法，$MS_{总}=SS_{总}/n_{总}$ ，其中 $SS_{总}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2$，$k$ 为处理组数，n_i 为第 i 组例数，$n_{总}=N-1$ 为总的自由度， N 表示总例数。

（二）方差分析的应用条件

（1） 各样本是相互独立的随机样本，且来自正态分布总体。

（2） 各样本的总体方差相等，即方差齐性(homoscedasticity)。

（三）不同设计资料的方差分析

1．完全随机设计的单因素方差分析

（1）资料类型：完全随机设计(completely random design)是将受试对象完全随机地分配到各个处理组。设计因素中只考虑一个处理因素，目的是比较各组平均值之间的差别是否由处理因素造成。

(2) 方差分析表：见表 5-1。$F\geq F_{\alpha}$ 时，拒绝 H_0: $m_1=m_2=\mathrm{L}\ \mathrm{L}=m_k$。

表 5-1 完全随机设计方差分析计算表

来源	SS	n	MS	F 值
组间	$SS_{组间}$	$n_{组间}=k-1$	$MS_{组间}=\frac{SS_{组间}}{n_{组间}}$	$F=\frac{MS_{组间}}{MS_{组内}}$
组内(误差)	$SS_{组内}=SS_{总}-SS_{组间}$	$n_{组内}=n_{总}-n_{组内}=N-k$	$MS_{组内}=\frac{SS_{组内}}{n_{组内}}$	
总计	$SS_{总}$	$n_{总}=N-1$		

2．随机区组设计的两因素方差分析

（1）资料类型：随机区组设计（randomized block design）是将受试对象按自然属性（如实验动物的窝别、体重，病人的性别、年龄及病情等）相同或相近者组成单位组（区组），然后把每个组中的受试对象随机地分配给不同处理。设计中有两个因素，一个是处理因素，另一个是按自然属性形成的单位组。单位组的选择原则是“单位组间差别越大越好，单位组内差别越小越好”。

（2）方差分析表：见表 5-2。$F_{处理}\geq F_{\alpha}$ 时，拒绝 H_0: $m_1=m_2=\mathrm{L}\ \mathrm{L}=m_k$。

表 5-2 随机区组设计方差分析计算表

变异来源	SS	n	MS	F 值
处理组间	$SS_{处理}$	$n_{处理}=k-1$	$MS_{处理}=\frac{SS_{处理}}{n_{处理}}$	$F_{处理}=\frac{MS_{处理}}{MS_{误差}}$

单位组间	$SS_{单位}$	$n_{单位}=b-1$	$MS_{单位}=\frac{SS_{单位}}{n_{单位}}$	$F_{单位}=\frac{MS_{单位}}{MS_{误差}}$
误差	$SS_{误差}=SS_{总}-SS_{处理}-SS_{单位}$	$n_{误差}=n_{总}-n_{处理}-n_{单位}=N-k-n+1$	$MS_{误差}=\frac{SS_{误差}}{n_{误差}}$	
总计	$SS_{总}$	$n_{总}=N-1$		

3．多个样本均数的多重比较

如果方差分析结果表明各组间有显著差别，则需要进一步进行两两比较，也称均数间的多重比较（multiple comparison）。进行两两比较的方法主要有：

（1） LSD-t 检验：称为最小显著差异 *t* 检验。适用于 *k* 组中某一对或某几对在专业上有特殊意义的均数间差异的比较。检验统计量为 *t* 值，自由度为方差分析表中的误差自由度，查 *t* 界值表。

$$t=\frac{\bar{X}_A-\bar{X}_B}{S_{\bar{d}_{AB}}} \qquad 其中\ S_{\bar{d}_{AB}}=\sqrt{MS_{误差}(\frac{1}{n_A}+\frac{1}{n_B})} \qquad (5\text{-}1)$$

（2） Dunnett-t 检验：它适用于 *k-1* 个试验组与一个对照组均数差别的多重比较，检验统计量为 *t* 值，自由度为方差分析表中的误差自由度，查 Dunnet-t 界值表。

$$t=\frac{|\bar{X}_i-\bar{X}_0|}{S_{\bar{X}_i-\bar{X}_0}} \text{，其中} S_{\bar{X}_i-\bar{X}_0}=\sqrt{MS_{误差}(\frac{1}{n_i}+\frac{1}{n_0})} \qquad (5\text{-}2)$$

（3）SNK-q 检验：在方差分析结果拒绝 H_0 时采用。适用于所有组均数的两两比较。检验统计量为 *q*，自由度为比较组数 *a* 和方差分析表中的误差自由度，查 *q* 界值表。

$$q=\frac{(\bar{X}_A-\bar{X}_B)}{S_{\bar{d}}} \qquad 其中，S_{\bar{d}}=\sqrt{\frac{MS_{误差}}{2}(\frac{1}{n_A}+\frac{1}{n_B})} \qquad (5\text{-}3)$$

4．多组资料方差起行检验

当各组标准差相差较大（如 1.5 倍）时，需检验资料是否满足方差齐性的条件。

5．变量变换

当资料不能满足方差分析的条件时，如果进行方差分析，可能造成错误的判断。因此对于明显偏离上述应用条件的资料，可以通过变量变换的方法来加以改善。常用的变量变换方法有：

（1）对数变换　对数变换不仅可以将对数正态分布的数据正态化，还能使数据方差达到齐性，特别是各样本的标准差与均数成比例或变异系数接近于一个常数时。变换公式为：

$$X'=\lg X \qquad (5\text{-}4)$$

当原始数据中有小值或零时，可用 $X'=\lg(X+1)$

（2）平方根变换　常用于使服从 Possion 分布的计数资料或轻度偏态的资料正态化；当各样本的方差与均数呈正相关时，可使资料达到方差齐性。变换公式为：

$$X'=\sqrt{X} \qquad (5\text{-}5)$$

当原始数据中有小值或零时，可用 $X'=\sqrt{X+0.5}$

（3）倒数变换　常用于数据两端波动较大的资料，可使极端值的影响减小。变换公式为：

$$X' = 1/X \qquad (5\text{-}6)$$

（4）平方根反正弦变换　常用于服从二项分布的率或百分比资料。一般地，当总体率较小（<30%）或较大（>70%）时，通过平方根反正弦变换，可使资料接近正态，且达到方差齐性的要求。变换公式为：

$$X' = \sin^{-1}\sqrt{X} \qquad (5\text{-}7)$$

（5）秩转换后，采用秩和检验比较组间差别（祥见第九章）。

6．两因素析因设计方差分析

处理含有两因素两水平的全面组合。例如治疗肿瘤术后病人，可采用 4 种方法：既不放疗也不化疗（a_0b_0）；放疗不化疗（a_1b_0）；不放疗化疗（a_0b_1）；既放疗又化疗（a_1b_1）。设放疗为 A 因素（两水平），化疗为 B 因素（两水平），则构成 2×2 析因设计，目的是分析 A 的主效应，B 的主效应及 AB 的交互作用。

7．重复测量资料的方差分析

受试对象随机分组后，多次测量某一观察指标，以比较处理效应在不同时间点有无变化。如试验组和对照组的轻度高血压病人入院前、治疗后 1 天、2 天、3 天、4 天的血压变化。设处理分组为 A 因素，重复测量的时间点为 B 因素，目的是分析 A 的主效应和 AB 的交互作用。

三、典型试题分析

1．完全随机设计资料的方差分析中，必然有（　　）

A．$SS_{组内}<SS_{组间}$　　B．$MS_{组间}<MS_{组内}$

C．$MS_{总}=MS_{组间}+MS_{组内}$　　D．$SS_{总}=SS_{组间}+SS_{组内}$

答案：D

[评析]　本题考点：方差分析过程中离均差平方和的分解、离均差平方和与均方的关系。

方差分析时总变异的来源有：组间变异和组内变异，总离均差平方和等于组间离均差平方和与组内离均差平方差之和，因此，等式 $SS_{总}=SS_{组间}+SS_{组内}$ 是成立的。离均差平方和除以自由度之后的均方就不再有等式关系，因此 C 选项不成立。A、B 选项不一定成立。D 选项为正确答案。

2．单因素方差分析中，当 $P<0.05$ 时，可认为（　）。

A．各样本均数都不相等　　B．各总体均数不等或不全相等

C．各总体均数都不相等　　D．各总体均数相等

答案：B

[评析] 本题考点：方差分析的检验假设及统计推断。

方差分析用于多个样本均数的比较，它的备择假设（H_1）是各总体均数不等或不全相等，当 $P<0.05$ 时，接受 H_1，即认为总体均数不等或不全相等。因此答案选 B。

3．以下说法中不正确的是（　）

A．方差除以其自由度就是均方

B．方差分析时要求各样本来自相互独立的正态总体

C．方差分析时要求各样本所在总体的方差相等

D．完全随机设计的方差分析时，组内均方就是误差均方

答案：A

[评析] 本题考点：方差分析的应用条件及均方的概念。

方差就是标准差的平方，也就是均方，因此选项 A 是错误的。选项 B、C 是方差分析对资料的要求，因此选项 B 和 C 都是正确的。在完全随机设计的方差分析中，组内均方就是误差均方，D 选项也是正确的。

4．当组数等于 2 时，对于同一资料，方差分析结果与 *t* 检验结果（　　）。

A.完全等价且 $F = t$　　B.方差分析结果更准确

C.*t* 检验结果更准确　　D.完全等价且 $t = \sqrt{F}$

答案：D

[评析]本题考点：方差分析与 *t* 检验的区别与联系。

对于同一资料，当处理组数为 2 时，*t* 检验和方差分析的结果一致且 $t = \sqrt{F}$，因此，正确答案为 D。

5．完全随机设计与随机单位组设计相比较（　）。

A.两种设计试验效率一样

B.随机单位组设计的误差一定小于完全随机设计

C.随机单位组设计的变异来源比完全随机设计分得更细

D.以上说法都不对

答案：C。

[评析]：本题考点：两种设计及其方差分析的区别。

两种设计不同，随机区组设计除处理因素外，还考虑了单位组因素。进行方差分析时，变异来源多分解出一项：单位组间变异。因此 C 选项为正确答案。

四、习　题

（五）　名词解释

1．均方　2．方差分析基本思想　3．总变异　4．组间变异　5．组内变异

6．完全随机设计　7．随机区组设计

（六）　单项选择题

1．两样本均数的比较，可用（　）。

A．方差分析　　B．*t* 检验

C．两者均可　　D．方差齐性检验

2．配伍组设计的方差分析中，$n_{配伍}$ 等于（　）。

A．$n_{总}-n_{误差}$　　B．$n_{总}-n_{处理}$

C．$n_{总}-n_{处理}+n_{误差}$　　D．$n_{总}-n_{处理}-n_{误差}$

3．在均数为 μ，标准差为 σ 的正态总体中随机抽样，$|\bar{X} - m| \geq$（　）的概率为 5%。

A．1.96σ　　B．$1.96 s_{\bar{x}}$　　C．$t_{0.05/2n} S$　　D．$t_{0.05/2n} s_{\bar{x}}$

4．当自由度（n_1, n_2）及显著性水准 α 都相同时，方差分析的界值比方差齐性检验的界值（　）。

A．大　　B．小　　C．相等　　D．不一定

5．方差分析中变量变换的目的是（　）。

A．方差齐性化　B．曲线直线化　C．变量正态化　D．以上都对

6．下面说法中不正确的是（　）。

A．方差分析可以用于两个样本均数的比较

B．完全随机设计更适合实验对象变异不太大的资料

C．在随机区组设计中，每一个区组内的例数都等于处理数

D．在随机区组设计中，区组内及区组间的差异都是越小越好

7．随机单位设计要求（　）。

A．单位组内个体差异小，单位组间差异大

B．单位组内没有个体差异，单位组间差异大

C．单位组内个体差异大，单位组间差异小

D．单位组内没有个体差异，单位组间差异小

8．完全随机设计方差分析的检验假设是（　）。

A．各对比组样本均数相等　B．各对比组总体均数相等

C．各对比组样本均数不相等　D．各对比组总体均数不相等

9．完全随机设计、随机区组设计的 *SS* 和及自由度各分解为几部分（　）。

A．2，2　B．2，3　C．2，4　D．3，3

10．配对 *t* 检验可用哪种设计类型的方差分析来替代（　）。

A．完全随机设计　B．随机区组设计

C．两种设计都可以　D．AB 都不行

（三）简答题

1．*t* 检验和方差分析的应用条件？

2．如何合理选择检验水准 α ？

3．以 *t* 检验为例，说明检验假设中 α 和 *P* 的区别。

（四）计算题

1．某湖水在不同季节氯化物含量测定值如表 5-3 所示。问不同季节氯化物含量有无差别？若有差别，进行 32 个水平的两两比较。

表 5-3　某湖水不同季节氯化物含量（*mg/L*）

春	夏	秋	冬	
22.6	19.1	18.9	19.0	
22.8	22.8	13.6	16.9	
21.0	24.5	17.2	17.6	
16.9	18.0	15.1	14.8	
20.0	15.2	16.6	13.1	
21.9	18.4	14.2	16.9	
21.5	20.1	16.7	16.2	
21.2	21.2	19.6	14.8	
$\sum X_{ij}$	167.9	159.3	131.9	129.3	588.40

n_i	8	8	8	8	32
$\overline{X}_i$	20.99	19.91	16.49	16.16	18.39
$\sum X^2_{ij}$	3548.51	3231.95	2206.27	2114.11	11100.84
s^2_i	3.53	8.56	4.51	3.47	

2．根据表 5-4 **资料说明大白鼠感染脊髓灰质炎病毒后，再做伤寒或百日咳接种是否影响生存日数？若结论为“有影响”，请做多重比较（与对照组比）。**

表 5-4 **各组大鼠接种后生存日数**

	伤寒	**百日咳**	**对照**	
	5	6	8	
	7	6	9	
	8	7	10	
	9	8	10	
	9	8	10	
	10	9	11	
	10	9	12	
	11	10	12	
	11	10	14	
	12	11	16	
$\sum X_{ij}$	92	84	112	288
n_i	10	10	10	30
$\overline{X}_i$	9.2	8.4	11.2	9.6
$\sum X^2_{ij}$	886	732	1306	2924
s^2_i	4.4	2.93	5.73	

3．有三种抗凝剂（A_1，A_2，A_3）对一标本作红细胞沉降速度（一小时值）测定，每种抗凝剂各作5次，问三种抗凝剂对红细胞沉降速度的测定有无差别？

A_1：15　11　13　12　14
A_2：13　16　14　17　15
A_3：13　15　16　14　12

4．**用** Dunnett-t **法检验下表中四个处理组均数与对照组的均数的差别。**

表 5-5　**家免脑损伤后大脑左半球组织水含量（%）**

试验分组	n	$\overline{X}_i$	S
对照（未损伤）	8	78.86	0.43
损伤后0.5小时	5	79.65	0.68
损伤后3小时	5	79.77	0.66
损伤后6小时	8	80.94	0.75
治疗组	9	79.61	0.66

5．将36只大白鼠按体重相近的原则配为12个单位组，各单位组的3只大白鼠随机地分配到三个饲料组。一个月后观察尿中氨基氮的排出量（mg）。经初步计算，$SS_{总}=162$，$SS_{饲料}=8$，$SS_{误差}=110$。试列出该实验数据的方差分析表。

6．将18名原发性血小板减少症患者按年龄相近的原则配为6个单位组，每个单位组中的3名患者随机分配到A、B、C三个治疗组中，治疗后的血小板升高见表5-6，问3种治疗方法的疗效有无差别？

表5-6　不同人用鹿茸草后血小板的升高值　（$10^4/mm^3$）

年龄组	A	B	C
1	3.8	6.3	8.0
2	4.6	6.3	11.9
3	7.6	10.2	14.1
4	8.6	9.2	14.7
5	6.4	8.1	13.0
6	6.2	6.9	13.4

7．某研究人员以0.3ml/kg剂量纯苯给大鼠皮下注射染毒，每周3次，经45天后，使实验动物白细胞总数下降至染毒前的50%左右，同时设置未染毒组。两组大鼠均按照是否给予升高白细胞药物分为给药组和不给药组，试验结果见下表，试作统计分析。

表5-7　试验效应指标（吞噬指数）数据

未染毒组		染毒组	
不给药	给药	不给药	给药
3.80	3.88	1.85	1.94
3.90	3.84	2.01	2.25
4.06	3.96	2.10	2.03
3.85	3.92	1.92	2.10
3.84	3.80	2.04	2.08

五、 习题答题要点

（九） 名词解释

1．均方：均方差（MS）或方差，是由离均差平方和被自由度相除而得。

2．方差分析：方差分析（analysis of variance，ANOVA）就是根据资料的设计类型，即变异的不同来源将全部观察值总的离均差平方和与自由度分解为两个或多个部分，除随机误差外，其余每个部分的变异可由某个因素的作用（或某几个因素的交互作用）加以解释。

通过各变异来源的均方与误差均方比值的大小，借助 *F* 分布作出统计推断，判断各因素对观测指标有无影响。

3．总变异：样本中全部实验单位差异称为总变异。其大小可以用全部观察值的均方（方差）表示。

4．组间变异：各处理组样本均数之间的差异，受处理因素的影响，这种变异称为组间变异，其大小可用组间均方表示。

5．组内变异： 各处理组内部观察值大小不等，这种变异称为组内变异，可用组内均方表示。

6．完全随机设计：只考虑一个处理因素，将全部受试对象随机分配到各处理组，然后观察实验效应，这种设计叫做完全随机设计。

7．随机区组设计：事先将全部受试对象按自然属性分为若干区组，原则是各区组内的受试对象的特征相同或相近，且受试对象数与处理因素的水平数相等。然后再将每个区组内的观察对象随机地分配到各处理组，这种设计叫做随机区组设计。

（十） 单项选择题

1.C 2.D 3.B 4.B 5.D 6.D 7.A 8.B 9.B 10.B

（三） 简答题

1． t 检验和方差分析均要求各样本来自相互独立的正态总体且各总体方差齐。

2．设置检验水准应根据研究目的，结合专业知识和研究设计要求，在未获得样本信息之前决定，而不应受到样本结果的影响。

3．以 t 检验为例，α 和 P 都是用 t 分布尾部面积大小表示，所不同的是：α 表示 I 型错误的概率，即 H_0 为真而被错误地拒绝的概率值。α 是在统计分析时，根据 I 型错误危害的大小，预先规定的，即规定统计结果为“接受 H_1”时的误判率的界限值为 α（即检验水准）。P 值是由实际样本得出的统计结果为“接受 H_1”时误判率。根据 P 与 α 的大小关系作出“不拒绝 H_0”或“拒绝 H_0”的统计推断。

（四）计算题

1．完全随机设计单因素芳差分析

解：H_0：4 个季节湖水中氯化物含量相等，即 $\mu_1=\mu_2=\mu_3=\mu_4$

H_1：4 个季节湖水中氯化物含量不等或不全相等。

$\alpha=0.05$

$$C=(\sum\sum X_{ij})^2/n=588.4^2/32=10819.205$$

$$SS_{总}=\sum\sum X^2{}_{ij}-C=11100.84-10819.205=281.635$$

$$SS_{组间}=\sum[(\sum X_{ij})^2/n_i]-C$$

$$=(167.9^2+159.3^2+131.9^2+129.3^2)/8-10819.205$$

$$=141.170$$

$$SS_{组内}=SS_{总}-SS_{组间}=281.635-141170=140.465$$

表 5-8 方差分析表

变异来源	*SS*	*n*	*MS*	*F*

总变异	281.635	31	47.057	9.380
组间变异	141.170	3	5.017	
组内变异	140.465	28		

查 F **界值表，** $F_{0.05,3,28}=2.95$**。因** $F>F_{0.05,3,28}$**所以** $P<0.05$**。按** $\alpha=0.05$ **水准，拒绝** H_0**，接受** H_1**，认为不同季节湖水中氯化物含量不同或不全相同。**

用 SNK-q **检验进行各组均数间两两比较。**

H_0**：任意两对比组的总体均数相等，** $\mu_A=\mu_B$

H_1**：** $\mu_A\neq\mu_B$

$\alpha=0.05$

表 5-9 **四个样本均数顺序排序**

组别	**春**	**夏**	**秋**	**冬**
$\overline{X}_i$	20.99	19. 91	16.49	16.16
位次	1	2	3	4

表 5-10 **四组均数两两比较** q **检验**

对比组	**两均数之差**	**组数**	q **值**	P **值**
1 , 4	4. 83	4	6. 099	<0.01
1 , 3	4. 50	3	5. 682	<0.01
1 , 2	1. 08	2	1.364	>0.05
2 , 4	3. 30	3	4. 735	<0.01
2 , 3	3. 42	2	4. 319	<0.01
3 , 4	0. 33	2	0. 417	>0.05

春与夏、秋与冬湖水中氯化物含量 $P>0.05$**，按** $\alpha=0.05$ **水准，不拒绝** H_0**，即不能认为春与夏、秋与冬季湖水中氯化物含量有差别。而其它** 4 **组均有** $P<0.01$**，按** $\alpha=0.05$ **水准，拒绝** H_0**，接受** H_1**，即认为春夏两季湖水中氯化物含量高于秋冬两季。**

2．完全随机设计单因素方差分析

H_0**：大白鼠感染脊髓灰质炎病毒后，再接种伤寒或百日咳菌苗生存日数相等．**

H_1**：大白鼠感染脊髓灰质炎病毒后，再接种伤寒或百日咳菌苗生存日数不等或不全相等．**

$\alpha=0.05$

$$C=(\sum\sum X_{ij})^2/n=288^2/30=2764.8$$

$$SS_{总}=\sum\sum X^2{}_{ij}-C=2924-2764.8=159.2$$

$$SS_{组间}=\sum[(\sum X_{ij})^2/n_i]-C$$

$$=(92^2+84^2+112^2)/10-2764.8=41.6$$

$SS_{组内} = SS_{总} - SS_{组间} = 159.2 - 41.6 = 117.6$

表 5-11　方差分析表

变异来源	SS	n	MS	F
总变异	159.2	29		
组间变异	41.6	2	20.80	4.77
组内变异	117.6	27	4.36	

查 F 界值表，$F_{0.05,2,27} = 3.35$。因 F 〉 $F_{0.05,2,27}$ 得 P <0.05，按 α=0.05 水准，拒绝 H_0，接受 H_1，认为大白鼠感染脊髓灰质炎病毒后，再接种伤寒或百日咳菌苗对生存日数有影响.

用 Dunnet -t 检验方法进行均数间多重比较：

H_0：任一组与对照组总体均数相同

H_1：任一组与对照组总体均数不同

α=0.05

由 Dunnett-t 检验公式，伤寒与对照组比较：

$$t = \frac{\bar{X}_1 - \bar{X}_0}{\sqrt{MS_{误差}(1/n_1 + 1/n_2)}}$$

$$= (9.2\text{-}11.2)/\sqrt{4.36(1/10 + 1/10)} = -2/0.93 = -2.14$$

n =27，查 Dunnett-t 检验界值表，得 P<0.05。按 α=0.05 水准，拒绝 H_0，接受 H_1，故可认为接种伤寒菌苗组较对照组生存日数减少。

百日咳与对照组比较：

$$t_{百对} = (8.4 - 11.2)/\sqrt{4.36(1/10 + 1/10)} = -2.99$$

n =27，查 Dunnett-t 检验界值表，得 P<0.05，按 α=0.05 水准，拒绝 H_0，接受 H_1，认为接种百日咳菌苗组较对照组生存日数减少。

3．完全随机设计资料方差分析

H_0：三种抗凝剂所作血沉值之间没有差别

H_1：三种抗凝剂所作血沉值之间存在差别

α=0.05

表 5-12　方差分析表

变异来源	SS	n	MS	F
总变异	40	14		
组间变异	10	2	5	2
组内变异	30	12	2.5	

查 F 界值表，$F_{0.05,2,12}$ =3.88 所以 P>0.05，按 α=0.05 水准，不能拒绝 H_0。即尚不能认为

三种抗凝剂所作血沉值之间有差别。

4．首先计算误差均方

$$SS_{误差}=\sum(n_i-1)s_i^2$$

$$=(8-1)\times0.43^2+(5-1)\times0.68^2+(5-1)\times0.66^2+(8-1)\times0.75^2+(9-1)\times0.66^2$$

$$=12.3086$$

$$n_{误差}=\mathrm{N}-\mathrm{k}=35-5=30$$

$$MS_{误差}=SS_{误差}/n_{误差}=12.3086/30=0.4103$$

（1） 损伤后0.5小时与对照组比

H_0：损伤后0.5小时与对照组组织含水量相等

H_1：损伤后0.5小时与对照组组织含水量不等

α=0.05

$$t=\frac{79.65-78.86}{\sqrt{0.4103(\frac{1}{8}+\frac{1}{5})}}$$

$$=2.16$$

以$n_{误差}=30$，处理数=4 查 Dunnett-t 界值表，得界值 2.25，因 t=2.16<2.25, 所以 P＞0.05。在 α=0.05 水准上，不拒绝H_0，尚不能认为损伤后0.5小时与对照组组织含水量有差别。

（2）损伤后3小时与对照组比

H_0：损伤后3小时与对照组组织含水量相等

H_1：损伤后3小时与对照组组织含水量不等

α=0.05

$$t=\frac{79.77-78.86}{\sqrt{0.4103(\frac{1}{8}+\frac{1}{5})}}=2.49$$

因 t>2.25（界值），故 P<0.05。在 α=0.05 水准上，拒绝 H_0，认为损伤后3小时与对照组的组织含水量有差别。

（3）损伤后6小时与对照组比

H_0：损伤后6小时与对照组组织含水量相等

H_1：损伤后6小时与对照组组织含水量不等

α=0.05

$$t=\frac{80.94-78.86}{\sqrt{0.4103(\frac{1}{8}+\frac{1}{8})}}=6.49$$

因 t>2.25（界值），故 P<0.05。在 α=0.05 水准上，拒绝 H_0，认为损伤后6小时与对照组的组织含水量有差别。

（4）治疗组与对照组比

H_0：治疗组与对照组组的织含水量相等

H_1：治疗组与对照组的组织含水量不等

α=0.05

$$t=\frac{79.61-78.86}{\sqrt{0.4103(\frac{1}{8}+\frac{1}{9})}}=2.41$$

因 t>2.25（界值），故 P<0.05。在 α=0.05 水准上，拒绝 H_0，认为治疗组与对照组的组织含水量有差别。

5．随机去组设计方差分析，总例数 N=36，处量组数 k=3，区组数 n=12。

计算： $SS_{区组} = SS_{总} - SS_{饲料} - SS_{误差} = 162 - 8 - 110 = 44$

$$\nu_{总} = N - 1 = 36 - 1 = 35$$

$$\nu_{饲料} = k - 1 = 3 - 1 = 2$$

$$\nu_{区组} = n - 1 = 12 - 1 = 11$$

$$\nu_{误差} = N - k - n + 1 = 36 - 3 - 12 + 1 = 22$$

根据计算结果填写方差分析表，见表 5-11**。**

表 5-13 **方差分析表**

变异来源	SS	n	MS	F	P
处理间	8	2	4	0.8	>0.05
区组间	44	11	4	0.8	>0.05
误差	110	22	5		
总变异	162	35			

6．**解：这两组资料用随机区组的方差分析为宜。**

（1**）处理组间比较**

H_0：**不同治疗组血小板升高值相同**

H_1：**不同治疗组血小板升高值不全相同**

α=0.05

（2**）年龄组间比较**

H_0：**不同年龄组血小板升高值相同**

H_1：**不同年龄组血小板升高值不全相同**

α=0.05

（3**）计算，列方差分析表**

表 5-14 **方差分析表**

变异来源	SS	n	MS	F
总变异	**187.265**	**17**		
组间	**129.003**	**2**	**64.502**	**79.338**
区组间	**50.132**	**5**	**10.026**	**12.333**
误差	**8.13**	**10**	**0.813**	

查 F **界值表，** $F_{0.05,2,10} = 4.10$, $F_{0.05,4,10} = 3.48$, **因此，组间及区组间均为** P<0.05**。按** α=0.05 **水准，拒绝** H_0**，可认为不同治疗组间血小板升高值不相同，不同年龄组患者血小板升高值也不相同。**

7．**设** A **因素为染毒（**2 **水平）**，B **因素为药物（**2 **水平），做** 2×2 **表析因设计方差分析。结果见表** 5-15**。**

表 5-15 **方差分析表**

变异来源	SS	n	MS	F

总变异	17.339	19		
染毒	0.009	1	0.009	1.000
药物	17.168	1	17.168	1907.555
染毒 * 药物	0.014	1	0.014	1.555
误差	0.148	16	0.009	

查 F 界值表，$F_{0.01,1,16}=8.68$，因此，药物组间 $P<0.01$。按 $\alpha=0.01$ 水准，认为给药组和不给药组吞噬指数不相同。

（赵清波　张玉海）

第六章　分类资料的统计描述

一、教学大纲要求

（一）掌握内容

1． 绝对数。

2． 相对数常用指标：率、构成比、比。

3． 应用相对数的注意事项。

4． 率的标准化和动态数列常用指标：标准化率、标准化法、时点动态数列、时期动态数列、绝对增长量、发展速度、增长速度、定基比、环比、平均发展速度和平均增长速度。

（二）熟悉内容

1． 标准化率的计算。

2． 动态数列及其分析指标。

二、教学内容精要

(一) 绝对数

绝对数是各分类结果的合计频数，反映总量和规模。如某地的人口数、发病人数、死亡人数等。绝对数通常不能相互比较，如两地人口数不等时，不能比较两地的发病人数，而应比较两地的发病率。

（二）常用相对数的意义及计算

相对数是两个有联系的指标之比，是分类变量常用的描述性统计指标，常用两个分类的绝对数之比表示相对数大小，如率、构成比、比等。

常用相对数的意义及计算见表6-1。

表6-1　常用相对数的意义及计算

常用相对数	概念	表示方式	计算公式	举例
率（rate）	又称频率指标，说明一定时期内某现象发生的频率或强度	百分率（%）、千分率（‰）等	率 $=\frac{\text{发生某现象的观察单位数}}{\text{可能发生某现象的观察单位总数}}\times 100\%$	单位时间内的发病率 患病率，如年（季）发病率、时点患病率等
构成比（proportion）	又称构成指标，说明某一事物内部各组成部分所占的比重或分布	百分数	构成比 $=\frac{\text{某一组成部分的观察单位数}}{\text{同一事物各组成部分的观察单位总数}}\times 100\%$	疾病或死亡的顺位、位次或所占比重
比	又称相对比，是 A、	倍数或分数	比 $=\frac{A}{B}$	①对比指标，如男：女

（ratio）	B 两个有关指标之比，说明 A 是 B 的若干倍或百分之几	=106.04∶100 ②关系指标，如医护人员：病床数=1.64 ③计划完成指标，如完成计划的 130.5%

(三) 应用相对数时应注意的问题

1．计算相对数的分母一般不宜过小。

2．分析时不能以构成比代替率 容易产生的错误有

（1）指标的选择错误如住院病人只能计算某病的病死率，不能认为是某病的死亡率；

（2）若用构成指标下频率指标的结论将导致错误结论，如 某部队医院收治胃炎的门诊人数中军人的构成比最高，但不一定军人的胃炎发病率最高。

3．不能用构成比的动态分析代替率的动态分析。

4．对观察单位数不等的几个率，不能直接相加求其总率。

5．在比较相对数时应注意可比性 通常应注意：

（1）观察对象，研究方法、观察时间、地区和民族等因素应相同或相近；

（2）其它影响因素在各组的内部构成是否相同。

6．对样本率（或样本构成比）的比较应随机抽样，并做假设检验。

（四）标准化法

1．标准化法（standardization method**）的意义和基本思想 常用于内部构成不同的两个或多个率的比较。标准化法的基本思想就是指定一个统一“标准”(标准人口构成比或标准人口数)，按指定“标准”计算调整率，使之具备可比性以后再比较，以消除由于内部构成不同对总率比较带来的影响。**

2．标准化率的计算 标准化率（standardized rate**）亦称调整率（**adjusted rate**）。常用的计算方法按已知条件有直接法和间接法。**

3．标准化法使用注意事项，如只用于组间比较，不能替代实际率等。

（五）动态数列及其分析指标

1．动态数列(dynamic series**)是一系列按时间顺序排列起来的统计指标，包括绝对数、相对数或平均数，用以说明事物在时间上的变化和发展趋势。**

2．动态数列依据时间上的特点可分为

⑴时点动态数列；

⑵时期动态数列。

3．动态数列常用的分析指标主要有

⑴绝对增长量；

⑵发展速度和增长速度，可计算

1） 定基比，即统一用某个时间的指标作基数，其它各时间的指标都与之相比；

2） 环比，即以前一个时间的指标作基数，以相邻的后一个时间的指标与之相比。

⑶平均发展速度和平均增长速度。

平均发展速度= $\sqrt[n]{a_n/a_0}$

平均增长速度= 平均发展速度-1

三、典型试题分析

（一）单项选择题

1．某医院某年住院病人中胃癌患者占4%，则（　　）。

A．4%是强度百分数　　B．4%是构成比

C．4%是相对比　　D．4%是绝对数

答案：B

[评析]　本题考点：对相对数概念的理解。

常用的相对数有率、构成比、比等。构成比又称构成指标，说明某是一事物内部各组成部分所占的比重或分布。胃癌患者是该年全部住院病人的一组成部分，占住院病人的 4%，则 4%是构成比。特别注意率与构成比的区别与联系，两者经常容易混淆。

2．欲比较两地死亡率，计算标准化率可以（　　）。

A．消除两地总人口数不同的影响

B．消除两地各年龄组死亡人数不同的影响

C．消除两地各年龄组人口数不同的影响

D．消除两地抽样误差不同的影响。

答案：C

[评析] 本题考点：标准化法的意义及应用。

标准化法常用于内部构成不同的两个或多个率的比较。标准化法的目的，就是为了消除由于内部构成不同对总率比较带来的影响，使调整以后的总率具有可比性。故欲比较两地死亡率，计算标准化率可以消除两地年龄别人口数不同对死亡率的影响。

3．计算麻疹疫苗接种后血清检查的阳转率，分母为（　　）。

A．麻疹易感人群　　B．麻疹患者数

C．麻疹疫苗接种人数　　D．麻疹疫苗接种后的阳转人数

答案：C

[评析] 本题考点：对相对数中率的概念的理解。

率又称频率指标，说明某现象发生的频率或强度。其公式为：

$$率=\frac{发生某现象的观察单位数}{可能发生某现象的观察单位总数}\times 100\%$$

，计算麻疹疫苗接种后血清检查的阳转率，分母为可能发生血清阳转的人数，即为麻疹疫苗接种人数。

（二）是非题

1.某医院收治某病患者 10 人，其中 8 人会吸烟，占 80%，则结论为“吸烟是发生该病的原因”。

答案：错。

[评析] 本题考点：对相对数概念的理解。

某医院收治某病患者 10 人，其中 8 人会吸烟，占 80%，则 80%为构成比或结构相对数。如果要探讨吸烟是否为发生该病的原因，应该比较吸烟人群与不吸烟人群该病的患病率。分析时不能以构成比代替率，若用构成指标下频率指标的结论将导致错误结论。

2.某化工厂某病连续 4 年患病率分别为 6.0%、9.7%、11.0%、15.4%，则该病 4 年总患病

率为：(6.0+9.7+11.0+15.4) /4=10.53 (%)。

答案：错。

[评析] 本题考点：对应用相对数时应注意的问题的理解。

应用相对数时对观察单位数不等的几个率，不能直接相加求其总率，而应该用总患病人数计算。因此该化工厂某病4年总患病率为10.53%是错误的。

四、习　题

（七）　单项选择题

11．某病患者120人，其中男性114人，女性6人，分别占95%与5%，则结论为（　　）。

A. 该病男性易得　　B. 该病女性易得

C. 该病男性、女性易患率相等　　D. 尚不能得出结论

12．甲县恶性肿瘤粗死亡率比乙县高，经标准化后甲县恶性肿瘤标化死亡率比乙县低，其原因最有可能是（　　）。

A. 甲县的诊断水平高

B. 甲县的肿瘤防治工作比乙县好

C. 甲县的老年人口在总人口中所占比例比乙县小

D. 甲县的老年人口在总人口中所占比例比乙县大

13．已知男性的钩虫感染率高于女性。今欲比较甲乙两乡居民的钩虫感染率，但甲乡人口女多于男，而乙乡男多于女，适当的比较方法是（　　）。

A. 分别进行比较

B. 两个率比较的χ^2检验

C. 不具备可比性，不能比较

D. 对性别进行标准化后再比较

14．经调查得知甲乙两地的冠心病粗死亡率为40/10万，按年龄构成标化后，甲地冠心病标化死亡率为45/10万；乙地为38/10万，因此可以认为（　　）。

A. 甲地年龄别人口构成较乙地年轻

B. 乙地年龄别人口构成较甲地年轻

C. 甲地冠心病的诊断较乙地准确

D. 甲地年轻人患冠心病较乙地多

15．某地区某种疾病在某年的发病人数为a_0，以后历年为a_1，a_2，……，a_n，则该疾病发病人数的年平均增长速度为（　　）。

A．$\frac{a_0+a_1+...+a_n}{n+1}$　　B．$\sqrt[n+1]{a_0\times a_1\times a_n}$

C．$\sqrt[n]{\frac{a_n}{a_0}}$　　D．$\sqrt[n]{\frac{a_n}{a_0}}-1$

16．某部队夏季拉练，发生中暑21例，其中北方籍战士为南方籍战士的2.5倍，则结论为（　　）。

A．　北方籍战士容易发生中暑

B．　南方籍战士容易发生中暑

C． 北方、南方籍战士都容易发生中暑

D． 尚不能得出结论

17．某地区某种疾病在某年的发病人数为 a_0，以后历年为 a_1，a_2，……，a_n，则该疾病发病人数的年平均发展速度为（ ）。

A．$\frac{a_0+a_1+...+a_n}{n+1}$　　B．$\sqrt[n+1]{a_0\times a_1\times a_n}$

C．$\sqrt[n]{\frac{a_n}{a_0}}$　　D．$\sqrt[n]{\frac{a_n}{a_0}}-1$

18．相对比包括的指标有（ ）。

A． 对比指标　　B．计划完成指标

C． 关系指标　　D．以上都是

（八） 名词解释

1．相对数　2．率　3．构成比　4．比　5．标准化法

6．动态数列　7．时点动态数列　8．定基比　9．环比　10.平均增长速度

（九） 简答题

1．常用的相对数指标有哪些？它们的意义和计算上有何不同？

2．为什么不能以构成比代率？请联系实际加以说明。

3．应用相对数时应注意哪些问题？

（十） 计算题

1. 某医院现有工作人员 900 人，其中男性 760 人，女性 140 人，在一次流感中发病者有 108 人，其中男性患者 79 人，而女性患者 29 人。试计算：

⑴该院总流感发病率？

⑵男、女流感发病率？

⑶男、女患者占总发病人数的百分比？

2. 下表为一抽样研究资料，试：填补空白处数据并根据最后三栏结果作简要分析。

表 6-2 某地各年龄组恶性肿瘤死亡情况

年龄（岁）	人口数	死亡总数	其中恶性肿瘤死亡数	恶性肿瘤死亡占总死亡的%	恶性肿瘤死亡率（1/10 万）	年龄别死亡率（‰）
⑴	⑵	⑶	⑷	⑸	⑹	⑺
0~	82920		4	2.90		
20~		63		19.05	25.73	
40~	28161	172	42			
60 及以上			32			
合计	167090	715	90	12.59		

3. 某城市 1971~1981 年乙脑发病率如下，试作动态分析。

表 6-3 某城市 1971~1981 年乙脑发病率（1/10 万）

年份	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981
发病率	20.52	6.31	1.87	3.07	1.08	1.38	2.29	2.31	2.47	2.76	2.94

4.试就下表资料分析比较甲、乙两医院乳腺癌手术后的五年生存率。

表 6-4 甲、乙两医院乳腺癌手术后的五年生存率（%）

腋下淋巴结转移	甲 医 院			乙 医 院		
	病例数	生存数	生存率	病例数	生存数	生存率
无	45	35	77.77	300	215	71.67
有	710	450	68.38	83	42	50.60
合计	755	485	64.24	383	257	67.10

五、习题答题要点

（十一）单项选择题

1.D 2.D 3.D 4.B 5.D 6.D 7.C 8.D

（十二）名词解释

1．相对数（relative number）是两个有联系的指标之比，是分类变量常用的描述性统计指标，常用相对数有率、构成比、比等。

2．率（rate）又称频率指标，说明一定时期内某现象发生的频率或强度。计算公式为：$率=\frac{发生某现象的观察单位数}{可能发生某现象的观察单位总数}\times100\%$，表示方式有：百分率（%）、千分率（‰）等。

3．构成比（proportion）又称构成指标，说明某一事物内部各组成部分所占的比重或分布。计算公式为：$构成比=\frac{某一组成部分的观察单位数}{同一事物各组成部分的观察单位总数}\times100\%$，表示方式有：百分数等。

4．比（ratio）又称相对比，是 A、B 两个有关指标之比，说明 A 是 B 的若干倍或百分之几。计算公式为：$比=\frac{A}{B}$，表示方式有：倍数或分数等。

5．标准化法(standardization method)是常用于内部构成不同的两个或多个率比较的一种方法。标准化法的基本思想就是指定一个统一“标准”(标准人口构成比或标准人口数)，按指定“标准”计算调整率，使之具备可比性以后再比较，以消除由于内部构成不同对总率比较带来的影响。

6．动态数列(dynamic series)是一系列按时间顺序排列起来的统计指标，包括绝对数、相对数或平均数，用以说明事物在时间上的变化和发展趋势。

7．时点动态数列是依据指标在时间方面的特点划分的一种动态数列,各个指标是在时点上的数据，如历年人口数、性别比例、现场调查中的患病人数、时点患病率等。

8．定基比即统一用某个时间的指标作基数，其它各时间的指标与之相比。

9．环比即以前一个时间的指标作基数，以相邻的后一个时间的指标与之相比。

10.平均增长速度是用于概括某一时期的平均速度变化，即该时期环比的几何均数减 1，其计算公式为：平均增长速度 = 平均发展速度 $-1=\sqrt[n]{a_n/a_0}-1$

（十三）简答题

1.常用的相对数指标有：率、构成比和相对比。意义和计算公式如下：

$$率=\frac{发生某现象的观察单位数}{可能发生某现象的观察单位总数}\times 100\%$$

率又称频率指标，说明某现象发生的频率或强度，常以 100%、1000‰等表示。

构成比又称构成指标，说明某一事物内部各组成部分所占的比重或分布。常以百分数表示。

$$构成比=\frac{某一组成部分的观察单\ 位数}{同一事物各组成部分的\ 观察单位总数}\times 100\%$$

比又称相对比，是 A、B 两个有关指标之比，说明两者的对比水平，常以倍数或百分数表示，其公式为：相对比=甲指标 / 乙指标（或 100%）

甲乙两个指标可以是绝对数、相对数或平均数等。

2.率和构成比所说明的问题不同，绝不能以构成比代率。构成比只能说明各组成部分的比重或分布，而不能说明某现象发生的频率或强度。例如：以男性各年龄组高血压分布为例，50~60 岁年龄组的高血压病例占 52.24%，所占比重最大，60~岁组则只占到 6.74%。这是因为 60~岁以上受检人数少，造成患病数低于 50~60 岁组，因而构成比相对较低。但不能认为年龄在 50~60 岁组的高血压患病率最严重，而 60 岁以上反而有所减轻。若要比较高血压的患病率，应该计算患病率指标。

3．应用相对数时应注意的问题有：

⑴ 计算相对数的分母一般不宜过小。

⑵ 分析时不能以构成比代替率。

⑶ 不能用构成比的动态分析代替率的动态分析。

⑷ 对观察单位数不等的几个率，不能直接相加求其总率。

⑸ 在比较相对数时应注意可比性。

⑹ 对样本率（或构成比）的比较应随机抽样，并做假设检验。

（十四）计算题：

1.⑴ 该院总流感发病率为：（108／900）×100%＝12%

⑵ 男性流感发病率为：（79／760）×100%＝10.39% ；

女性流感发病率为：（29／140）×100%＝20.71%

⑶ 男性患者占总发病人数的百分比为：（79／108）×100%＝73.15% ；

女性患者占总发病人数的百分比为：（29／108）×100%＝26.85%

2.填补空白处数据，见下表（ ）内。

表 6-5 某地各年龄组恶性肿瘤死亡情况

年龄（岁）	人口数	死亡总数	其中恶性肿瘤死亡数	恶性肿瘤死亡占总死亡的%	恶性肿瘤死亡率（1/10 万）	年龄别死亡率（‰）
⑴	⑵	⑶	⑷	⑸=⑷/⑶	⑹=⑷/⑵	⑺

0~	82920	(138)	4	2.90	(4.82)	(1.66)
20~	(46638)	63	(12)	19.05	25.73	(1.35)
40~	28161	172	42	(24.42)	(149.14)	(6.11)
60~	(9371)	(342)	32	(9.36)	(341.48)	(36.50)
合计	167090	715	90	12.59	(53.86)	(4.28)

根据最后三栏结果作简要分析。

由表中第(5)栏可知：40~**岁组恶性肿瘤死亡占总死亡比重最高，近** 1/4**；**20~**岁组次之，占** 19.05%**；**60~**岁组恶性肿瘤死亡人数虽多，但仅占总死亡的** 9.36%**；**0~**岁组恶性肿瘤死亡占总死亡比重最低，仅占**2.90%**。**

由表中第(6)栏可知：恶性肿瘤的年龄别死亡率随年龄的增大而增加，以60~**岁组为最高，为** 341.50/10 **万。故可认为随年龄增大，患恶性肿瘤的危险增加，应引起足够的重视。**

由表中第(7)栏可知：年龄别死亡率以20 **至**40 **岁最低，以后随年龄的增加而增加，**60 **岁以后高达**36.50‰**。**

3**.计算结果见表** 6-6**。**

表 6-6　**某市** 1971~1981 **年乙脑发病率动态分析**

年份	发病率	绝对增长量		发展速度（%）		增长速度（%）	
	（1/10 万）	累计	逐年	定基比	环比	定基比	环比
1971	20.52	—	—	100	100	—	—
1972	6.31	-14.21	-14.21	30.75	30.75	-69.25	-69.25
1973	1.87	-18.56	-4.44	9.11	29.64	-90.89	-70.36
1974	3.07	-17.45	1.20	14.96	164.17	-85.04	64.17
1975	1.08	-19.44	-1.99	5.26	35.18	-94.74	-64.82
1976	1.38	-19.14	0.30	6.73	127.78	-93.27	27.78
1977	2.29	-18.23	0.91	11.16	165.94	-88.84	65.94
1978	2.31	-18.21	0.02	11.26	100.87	-88.74	0.87
1979	2.47	-18.05	0.16	12.04	106.93	-87.96	6.93
1980	2.76	-17.76	0.29	13.45	111.74	-86.55	11.74
1981	2.94	-17.58	0.18	14.33	106.52	-85.67	6.52

4**.两医院乳腺癌患者的病情构成不同，比较两医院的标准化率，计算过程见表** 6-7**。**

表 6-7 **甲、乙两医院乳腺癌手术后的五年生存率标化（甲乙两医院合计为标准）**

腋下淋巴结转移	标准病例数 N_i	甲　医　院		乙　医　院	
		原生存率 P_i	预期生存人数 N_iP_i	原生存率 P_i	预期生存人数 N_iP_i
(1)	(2)	(3)	(4)=(2)(3)	(5)	(6)=(2)(5)
无	**345**	77.77	**268**	71.67	**247**
有	**793**	68.38	**503**	50.60	**401**

合计	**1138（Σ N_i）**	64.24	**771（Σ N_iP_i）**	67.10	**648（Σ N_iP_i）**

甲医院乳腺癌手术后的五年生存率标化生存率：

$$p' = \frac{\sum N_i P_i}{N} \times 100\% = \frac{771}{1138} \times 100\% = 67.75\%$$

乙医院乳腺癌手术后的五年生存率标化生存率：

$$p' = \frac{\sum N_i P_i}{N} \times 100\% = \frac{648}{1138} \times 100\% = 56.94\%$$

因为甲、乙两医院有无腋下淋巴结转移的病情构成不同，故标化后，甲医院乳腺癌手术后的五年生存率高于乙医院，校正了标化前甲医院低于乙医院的情况。

（蒋知俭　万毅）

第七章　二项分布与 Poisson 分布及其应用

一、教学大纲要求

（一）掌握内容

1．二项分布

（1）分布参数；

（2）各项统计指标（均数、标准差等）的计算方法；

（3）二项分布的分布特征，近似分布及其应用条件。

2．Poisson 分布

（1）分布参数；

（2）各项统计指标（均数、标准差等）的计算方法；

（3）Poisson 分布的分布特征，近似分布及其应用条件。

（二）熟悉内容

1．二项分布

（1）样本率的分布；

（2）总体率的区间估计；

（3）样本率与总体率的比较；

（4）两样本率的比较。

2．Poisson 分布

（1）总体均数的区间估计；

（2）样本均数与总体均数的比较；

（3）两个样本均数的比较。

（三）了解内容

二项分布及 Poisson 分布的前提条件及其概率密度函数的应用。

二、教学内容精要

（一）基本概念

1．概率分布

二项分布（binomial distribution）和 Poisson 分布是统计学中很重要的两种分布。

二项分布：若一个随机变量 X，它的可能取值是 $0,1,\ldots,n$，且相应的取值概率为

$$P(X=k)=\binom{n}{k}p^{k}(1-p)^{n-k} \tag{7-1}$$

则称此随机变量 X 服从以 n、π 为参数的二项分布，记为 $X\sim B(n,\pi)$。

Poisson 分布：若离散型随机变量 X 的取值为 $0,1,\ldots,n$，且相应的取值概率为

$$P(X=k)=\frac{m^{k}}{k!}e^{-m}\quad(\mu>0) \tag{7-2}$$

则称随机变量 X 服从以 μ 为参数的 Poisson 分布（Poisson Distribution），记为 $X\sim P(\mu)$。

2．两种分布成立的条件

（1）二项分布成立的条件：①每次试验只能是互斥的两个结果之一；②每次试验的条件不变；③各次试验独立。

（2）Poisson 分布成立的条件：①平稳性：X 的取值与观察单位的位置无关，只与观察单位的大小有关；②独立增量性：在某个观察单位上 X 的取值与前面各观察单位上 X 的取值无关；③普通性：在充分小的观察单位上 X 的取值最多为 1。

（二）分布参数

1．二项分布，$X\sim B(n,\pi)$

X 的均数 $\mu_X = n\pi$ (7-3)

X 的方差 $\sigma_X^2 = n\pi(1-\pi)$ (7-4)

X 的标准差 $\sigma_X = \sqrt{np(1-p)}$ (7-5)

2．Poisson 分布，$X\sim P(\mu)$

X 的均数 $\mu_X = \mu$ (7-6)

X 的方差 $\sigma_X^2 = \mu$ (7-7)

X 的标准差 $\sigma_X = \sqrt{m}$ (7-8)

（三）分布特性

1．可加性

二项分布和 Poisson 分布都具有可加性。

如果 X_1，X_2，… X_k 相互独立，且它们分别服从以 n_i，p（i=1,2, …,k）为参数的二项分布，则 $X=X_1+X_2+\cdots+X_k$ 服从以 n，p（$n=n_1+n_2+\cdots+n_k$）为参数的二项分布。如果 X_1，X_2，…，X_k 相互独立，且它们分别服从以 μ_i（i=1,2, …,k）为参数的 Poisson 分布，则 $X=X_1+X_2+\cdots+X_k$ 服从以 μ（$\mu=\mu_1+\mu_2+\cdots+\mu_k$）为参数的 Poisson 分布。

2．近似分布

特定条件下，二项分布、Poisson 分布可近似于某种其它的分布，这一特性拓宽了它们的应用范围。

二项分布的正态近似：当 n 较大，π 不接近 0 也不接近 1 时，二项分布 $B(n,\pi)$ 近似正态分布 $N(n\pi,\sqrt{np(1-p)})$。

二项分布的 Poisson 分布近似：当 n 很大，π 很小，$np=l$ 为一常数时，二项分布近似于 Poisson 分布。

Poisson 分布的正态近似：Poisson 分布 $P(\mu)$，当 μ 相当大时（≥20），其分布近似于正态分布。

（四）应用

1．二项分布的应用

（1） 总体率的区间估计

有查表法和正态近似法两种方法。

当 $n\leq 50$ 时可以通过查表求总体率的 95%和 99%可信区间。

当二项分布满足近似正态分布的条件时（n 较大，样本率 p 不接近 0 也不接近 1），可用正态近似法求总体率的 $1-\alpha$ 可信区间：

$$(p-u_\alpha S_p,\ p+u_\alpha S_p) \quad (7\text{-}9)$$

$$S_p=\sqrt{\frac{p(1-p)}{n}} \quad (7\text{-}10)$$

（2） 样本率与总体率比较

应用二项分布的概率计算公式计算事件（一般指X取某给定值一侧的所有值）发生的概率，再比较其与检验水准α大小，推断样本所在的总体率与给定总体率的关系。

（3） 两样本率的比较

根据独立的两个正态变量的差也服从正态分布的性质和二项分布在一定条件下的近似正态分布特性，当两个样本的含量n_1和n_2较大，且p_1、（1-p_1）、p_2、（1-p_2）均不太小，可用u检验方法对两样本率对应的总体率作统计推断。

$$u=\frac{p_1-p_2}{S_{p_1-p_2}} \tag{7-11}$$

$$S_{p_1-p_2}=\sqrt{\frac{X_1+X_2}{n_1+n_2}(1-\frac{X_1+X_2}{n_1+n_2})(\frac{1}{n_1}+\frac{1}{n_2})} \tag{7-12}$$

2．Poisson 分布的应用

（1） 总体均数的区间估计

有查表法和正态近似法两种方法。

当样本计数$X\leq 50$时，可用查表法求得总体均数的95%或99%可信区间。

当样本计数$X>50$时，可利用Poisson分布的正态近似性，计算其总体均数（1-α）可信区间如下：

$$(X-u_a\sqrt{X}\ ,\ X+u_a\sqrt{X}\ ) \tag{7-13}$$

（2）样本均数与总体均数的比较

有直接计算概率法和正态近似法两种方法。

样本均数与总体均数比较的目的是推断此样本所代表的未知总体均数μ是否等于已知总体均数μ_0。

当总体均数较小时，可采用直接计算概率法进行比较。X取某一值的概率以Poisson分布的概率密度函数来计算，即

$$P(X=k)=\frac{m^k}{k!}e^{-m}\quad (k=0,1,2,\ldots)$$

注意：样本均数与总体均数比较时，应以X取大于等于（样本均数大于总体均数时）或小于等于（样本均数小于总体均数时）样本均数的所有值的概率总和同检验界值α进行比较，切不可仅以X取样本均数的概率同检验界值进行比较。

当总体均数较大时，可用正态近似法进行统计推断。此时 Poisson 分布近似正态分布，故可计算标准正态统计量u，

$$u=\frac{X-u_0}{\sqrt{u_0}} \tag{7-14}$$

通过u值得出相应的概率，推断样本均数与总体均数的关系。

（3）两个样本均数的比较：两个样本计数均较大时，可根据Poisson分布的正态近似性对其进行u检验。

两个样本观察单位相同时，用下式计算u值。

$$u=\frac{X_1-X_2}{\sqrt{X_1+X_2}} \tag{7-15}$$

两个样本观察单位不同时，用下式计算 u 值。

$$u=\frac{X_1/n_1-X_2/n_2}{\sqrt{\frac{X_1}{n_1^2}+\frac{X_2}{n_2^2}}} \tag{7-16}$$

三、典型试题分析

(一)单项选择题

1．某地人群中高血压的患病率为 π，由该地区随机抽查 n 人，则（　）

A．样本患病率 $p=X/n$ 服从 B（n, π）

B．n 人中患高血压的人数 X 服从 B（n, π）

C．患病人数与样本患病率均不服从 B（n, π）

D．患病人数与样本患病率均服从 B（n, π）

答案：B

[评析]　本题考点：二项分布概念的理解。

二项分布中所指的随机变量 X 代表 n 次试验中出现某种结果的次数，具体到本题目就是指抽查的 n 个人中患高血压的人数，因此答案为 B。

2．二项分布近似正态分布的条件是（　）

A．n 较大且 π 接近 0　　B．n 较大且 π 接近 1

C．n 较大且 π 接近 0 或 1　　D．n 较大且 π 接近 0.5

答案：D

[评析] 本题考点：二项分布的正态近似特性。

从对二项分布特性的描述中可知：当 n 较大，π 不接近 0 也不接近 1 时，二项分布 B（n, π）近似正态分布 N（$n\pi$，$\sqrt{np(1-p)}$）。π 不接近 0 也不接近 1，等同于 π 接近 0.5，因而此题目答案为 D。

3．以下分布中，其均数和方差总是相等的是（　）

A．正态分布　　B．对称分布

C．Poisson 分布　　D．二项分布

答案：C

[评析] 本题考点：Poisson 分布的特性。

Poisson 分布 P（μ）的参数只有一个，即 μ。它的均数和方差均等于 μ，这一点大家需要牢记。

4．测得某地区井水中细菌含量为 10000／L,据此估计该地区每毫升井水中细菌平均含量的 95%可信区间为（　）

A．$10000 \pm 1.96\sqrt{10000}$　　B．$10 \pm 1.96\sqrt{10}$

C．$10 \pm 1.96\dfrac{\sqrt{10000}}{1000}$　　D．$10 \pm 1.96\sqrt{10000}$

答案：C

[评析]本题考点：Poisson 分布的正态近似性。

当 X 较大（一般大于 50）时，Poisson 分布近似正态分布，按照正态分布资料的计算公式计算该地区井水中平均每升细菌含量的 95%可信区间，再除以 1000 即得平均每毫升井水中细菌的平均含量（设 $Y = X/1000$，有 $S_Y = S_X/1000 = \sqrt{10000}/1000$）。

（二） 是非题

从装有红、绿、蓝三种颜色的乒乓球各 500、300、200 只的暗箱中随机取出 10 个球，以 X 代表所取出球中的红色球数，则 X 服从二项分布 B（10，0.5)。（ ）

答案：正确。

[评析] 本题考点：二项分布的定义。

二项分布成立的条件是：①每次试验只能是互斥的两个结果之一；②每次试验的条件不变；③各次试验独立。此题目所述情况完全满足后两个条件，关键在于第一个条件的判断，从表面上看，每次试验的结果有三种，但本题目所关心的试验结果是“红色与否”，因而该试验结果仍为两种互斥的情况—“红色”和“非红色”。所以，此题目所述情况满足以上三个条件，X 服从二项分布 B（10，0.5)。

（三）计算题

炮击命中目标的概率为 0.2，共发射了 14 发炮弹。已知至少要两发炮弹命中目标才能摧毁之，试求摧毁目标的概率。

答案：0.802

[评析] 本题的考点：二项分布概率函数的理解和应用能力。

摧毁目标的概率即有两发或两发以上炮弹命中目标的概率，此概率又等于 1 减去只有一发命中或无一命中的概率之差。根据二项分布的概率函数计算如下：

$$P_{X\geq 2} = 1 - P_{X\leq 1} = 1 - \left[(1-0.2)^{14} + \binom{14}{1} \times 0.2^1 \times (1-0.2)^{13}\right] = 1 - \left[0.044 + 0.154\right] = 0.802$$

四、习　题

（一）名词解释

1．二项分布　　2．Poisson 分布

3．Bernoulli 试验

（二）　单项选择题：

1. X_1、X_2 分别服从二项分布 B（n_1，p_1)、B（n_2，p_2)，且 X_1、X_2 相互独立，若要 $X = X_1 + X_2$ 也服从二项分布，则需满足下列条件（　）。

A．$X_1 = X_2$　　B. $n_1 = n_2$

C．$p_1 = p_2$　　D. $n_1 p_1 = n_2 p_2$

2. 二项分布 B（n，p）的概率分布图在下列哪种条件下为对称分布（　　）。

A．n=50　　B. p=0.5

C．np=1　　D. p=1

3. Poisson 分布 P（m）满足下列何种条件时近似正态分布 N（m，$\sqrt{m}$）（　　）。

A．m 相当大　　B. m =1

C．m =0　　D. m =0.5

4. 已知某高校学生近视眼的患病率为 50%，从该高校随机挑选 3 名学生，其中 2 人患近视眼的概率为（　　）。

A．0.125　　B. 0.375

C．0.25　　D. 0.5

5. 某自然保护区狮子的平均密度为每平方公里 100 只，随机抽查其中一平方公里范围内狮子的数量，若进行 100 次这样的抽查，其中的 95 次所得数据应在以下范围内（　　）。

A．5～195　　B. 80.4～119.6

C．95～105　　D．74.2～125.8

（三）简答题

1．服从二项分布及 Poisson 分布的条件分别是什么？

2．二项分布、Poisson 分布分别在何种条件下近似正态分布？

3．在何种情况下，可以用率的标准误 S_p 描述率的抽样误差？

（四）计算题

1. 已知我国成人乙肝病毒表面抗原平均阳性率为 10%，现随机抽查某地区 10 位成人的血清，其中 3 人为阳性。该地区成人乙肝表面抗原阳性率是否高于全国平均水平？

2. 对甲、乙两种降压药进行临床疗效评价，将某时间段内入院的高血压病人随机分为两组，每组均为 100 人。甲药治疗组 80 位患者有效，乙药治疗组 50 位患者有效，两种降压药有效率有无差别？

3. 某放射性物质发生脉冲频率为 100／克／小时，已知某矿区矿石中该放射性物质的含量为 4 克／千克，今又测得另一矿区同种矿石每千克发生脉冲频率为 1000／小时，问两个矿区矿石中该放射性物质的含量是否相等？

4. 一台仪器在 10000 个工作时内平均发生 10 次故障，试求在 100 个工作时内故障不多于两次的概率。

五、习题答题要点

（一）名词解释

1. 二项分布：若一个随机变量 X，它的可能取值是 0,1,…,n，且相应的取值概率为

$$P(X=k)=\binom{n}{k}p^{k}(1-p)^{n-k}$$

则称此随机变量 X 服从以 n、π 为参数的二项分布（Binomial Distribution），记为 $X\sim B(n,\pi)$。

2. Poisson 分布：若离散型随机变量 X 的取值为 0,1,…,n，且相应的取值概率为

$$P(X=k)=\frac{m^k}{k!}e^{-m} \quad (\mu>0)$$

则称随机变量 X 服从以 μ 为参数的 Poisson 分布（Poisson Distribution），记为 $X\sim P(\mu)$。

3. Bernoulli 试验：将感兴趣的事件 A 出现的试验结果称为“成功”，事件 A 不出现的试验结果称为“失败”，这类试验就称为 Bernoulli 试验（Bernoulli Test）。

（二）单项选择题

1.C　2.B　3.A　4.B　5.B

（三）问答题

1.二项分布成立的条件：①每次试验只能是互斥的两个结果之一；②每次试验的条件不变；③各次试验独立。Poisson 分布成立的条件：①平稳性：X 的取值与观察单位的位置无关，只与观察单位的大小有关；②独立增量性：在某个观察单位上 X 的取值与前面各观察单位上 X 的取值无关；③普通性：在充分小的观察单位上 X 的取值最多为 1。

2. 二项分布的正态近似：当 n 较大，π 不接近 0 也不接近 1 时，二项分布 B（n, π）近似正态分布 $N(n\pi, \sqrt{np(1-p)})$。

Poisson 分布的正态近似：Poisson 分布 $P(\mu)$，当 μ 相当大时（≥20），其分布近似于正态分布。

3. 当率 P 所来自的样本近似服从正态分布时，即 n 较大，P 不接近 0 也不接近 1 时，可以用率的标准误 S_p 描述率的抽样误差。

（四）计算题

1. 建立检验假设

H_0：该地区成人乙肝表面抗原阳性率为 10%；

H_1：该地区成人乙肝表面抗原阳性率大于 10%。

α=0.05。

从总体率为 10%的人群随机抽取 10 人，3 人或 3 人以上阳性的概率为：

$P(X\geq 3)=1-[P(X=0)+P(X=1)+P(X=2)]=1-[0.9^{10}+10*0.1*0.9^{9}+45*0.1^{2}*0.9^{8}]=0.0702$

$P(X\geq 3)>0.05$，在 α=0.05 水平上，不拒绝 H_0，不能认为该地区成人乙肝表面抗原阳性率高于全国水平。

2. 建立检验假设

H_0：两种药有效率无差别；

H_1：两种药有效率有差别。

α=0.05。

$$S_{p_1-p_2}=\sqrt{\frac{X_1+X_2}{n_1+n_2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$S_{p_1-p_2}=\sqrt{\frac{80+50}{100+100}\left(\frac{1}{100}+\frac{1}{100}\right)}=0.1095$$

$$u=\frac{p_1-p_2}{S_{p_1-p_2}}$$

$$u=\frac{0.8-0.5}{0.114}=2.6312>2.58，P<0.01$$

在 α=0.05 **水平上，拒绝** H_0**，接受** H_1**，即两种降压药有效率有显著差别，甲药比乙药有效率高。**

3. **放射性物质含量为** 4 **克／千克的矿石每千克的平均脉冲记数为** m =100*4=400**／小时，** m **值较大，可利用** Poisson **分布的近似正态分布特性进行计算。**

H_0**：两矿区矿石中该放射性物质含量相等，即后一矿区矿石发生脉冲频率的总体均数为** 400**／小时；** H_1**：两矿区矿石中该放射性物质含量不相等，即后一矿区矿石发生脉冲频率的总体均数不等于** 400**／小时。** α=0.05**。**

$$u=\frac{X-u_0}{\sqrt{u_0}}$$

$$u=\frac{1000-400}{\sqrt{400}}=30>2.58，P<0.01。$$

在 α=0.05 **水平上，拒绝** H_0**，接受** H_1**，即两矿区矿石中该放射性物质含量不相等，后一矿区矿石中该放射性物质含量高于前一矿区。**

4. **该仪器在** 100 **个工作时内故障不多于两次的概率即为** $P(X=0)$ **，** $P(X=1)$ **，** $P(X=2)$ **三者之和。而** 100 **个工作时内故障平均次数为** $m=100\times\frac{10}{10000}=0.1$ **，根据** Poisson **分布的概率函数计算如下：**

$$P(X\le 2)=\frac{m^0}{0!}e^{-m}+\frac{m^1}{1!}e^{-m}+\frac{m^2}{2!}e^{-m}=0.90484+0.09048+0.00452=0.99984$$

故该仪器在 100 **个工作时内故障不多于两次的概率为** 0.99984**。**

（夏结来　薛富波）

c^2 检验

一、教学大纲要求

（一） 掌握内容

1. c^2 检验的用途。

2. 四格表的 c^2 检验。

（1） 四格表 c^2 检验公式的应用条件；

（2） 不满足应用条件时的解决办法；

（3） 配对四格表的 c^2 检验。

3. 行×列表的 c^2 检验。

（二） 熟悉内容

频数分布拟合优度的 c^2 检验。

（三） 了解内容

1． c^2 分布的图形。

2．四格表的确切概率法。

二、教学内容精要

(一) c^2 检验的用途

c^2 检验（Chi-square test）用途较广，主要用途如下：

1．推断两个率及多个总体率或总体构成比之间有无差别

2．两种属性或两个变量之间有无关联性

3．频数分布的拟合优度检验

(二) c^2 检验的基本思想

1． c^2 检验的基本思想是以 c^2 值的大小来反映理论频数与实际频数的吻合程度。在零假设 H_0（比如 H_0：$p_1=p_2$）成立的条件下，实际频数与理论频数相差不应该很大，即 c^2 值不应该很大，若实际计算出的 c^2 值较大，超过了设定的检验水准所对应的界值，则有理由怀疑 H_0 的真实性，从而拒绝 H_0，接受 H_1（比如 H_1：$p_1\neq p_2$）。

2． 基本公式：$c^2=\sum\frac{(A-T)^2}{T}$，$A$ 为实际频数（Actual Frequency），T 为理论频数（Theoretical Frequency）。四格表 c^2 检验的专用公式正是由此公式推导出来的，用专用公式与用基本公式计算出的 c^2 值是一致的。

(三)率的抽样误差与可信区间

1．率的抽样误差与标准误

样本率与总体率之间存在抽样误差，其度量方法：

$s_p=\sqrt{\frac{p(1-p)}{n}}$，$p$ 为总体率，或 (8-1)

$S_p=\sqrt{\frac{p(1-p)}{n}}$， p 为样本率； (8-2)

2．总体率的可信区间

当 n 足够大，且 p 和 1-p 均不太小，p 的抽样分布逼近正态分布。

总体率的可信区间：（$p-u_{a/2}\times S_p, p+u_{a/2}\times S_p$）。 (8-3)

(四) c^2 检验的基本计算

见表 8-1。

表 8-1 c^2 检验的用途、假设的设立及基本计算公式

资料形式	用途	H_0、H_1 的设立与计算公式	自由度
四格表	①独立资料两样本率的比较	H_0：两总体率相等 H_1：两总体率不等 ①专用公式	1
	②配对资料两样本率的比较	$c^2=\frac{(ad-bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}$ ②当 n≥40 但 1≤T<5 时，校正公式 $c^2=\frac{(\lvert ad-bc\rvert-n/2)^2 n}{(a+b)(c+d)(a+c)(b+d)}$ ③配对设计 $c^2=\frac{(\lvert b-c\rvert-1)^2}{b+c}$	
R×C 表	①多个样本率、构成比的比较	H_0：多个总体率（构成比）相等 (H_0：两种属性间存在关联)	(R-1)(C-1)
	②两个变量之间关联性分析	H_1：多个总体率（构成比）不全相等 (H_0：两种属性间存在关联) $c^2=n(\sum\frac{A^2}{n_R n_C}-1)$	
频数分布表	频数分布的拟合优度检验	H_0：资料服从某已知的理论分布 H_1：资料不服从某已知的理论分布 $\sum\frac{(A-T)^2}{T}$	据频数表的组数而定

（五）四格表的确切概率法

当四格表有理论数小于 1 或 $n<40$ 时，宜用四格表的确切概率法。

（六）c^2 检验的应用条件及注意事项

1．分析四格表资料时，应注意连续性校正的问题，当 $1<T<5$，$n>40$ 时，用连续性校正 c^2 检验；$T\le 1$，或 $n\le 40$ 时，用 Fisher 精确概率法。

2．对于 $R\times C$ 表资料应注意以下两点：

（1）理论频数不宜太小，一般要求：理论频数<5 的格子数不应超过全部格子的 1/5；

（2）注意考察是否有有序变量存在。对于单向有序 $R\times C$ 表资料，当指标分组变量是有序的时，宜用秩和检验；对于双向有序且属性不同的 $R\times C$ 表资料，若希望弄清两有序变量之间是否存在线性相关关系或存在线性变化趋势，应选用定性资料的相关分析或线性趋势检验；对于双向有序且属性相同的 $R\times C$ 表资料，为考察两种方法检测的一致性，应选用 Kappa 检验。

三、典型试题分析

（一）单项选择题

1．下列哪项检验不适用 c^2 检验（　　）

A． 两样本均数的比较

B． 两样本率的比较

C． 多个样本构成比的比较

D． 拟合优度检验

答案：A

[评析] 本题考点：χ^2 检验的主要用途。χ^2 检验不能用于均数差别的比较。

2．分析四格表时，通常在什么情况下需用 Fisher 精确概率法（　　）

A．$1<T<5, n>40$　　B．$T<5$　　C．$T\leq 1$ 或 $n\leq 40$　　D．$T\leq 1$ 或 $n\leq 100$

答案： C

[评析] 本题考点：对于四格表，当 $T\leq 1$ 或 $n\leq 40$ 时，不宜用 χ^2 检验，应用 Fisher 精确概率法。

3．χ^2 值的取值范围为

A．$-\infty < \chi^2 < +\infty$　　B．$0 \leq \chi^2 \leq +\infty$　　C．$\chi^2 \leq 1$　　D．$-\infty \leq \chi^2 \leq 0$

答案： B

[评析] 根据 χ^2 分布的图形或 χ^2 的基本公式可以判断 χ^2 值一定是大于等于零且没有上界的，故应选 B。

（二）是非题

两样本率的比较可以采用 χ^2 检验，也可以采用双侧 u 检验。

答案：正确。

[评析] 就两个样本率的比较而言，双侧 u 检验与 χ^2 检验是等价的。

（三）简答题

1．四格表的 χ^2 检验和 u 检验有何联系与区别？

答案：相同点：凡是能用 u 检验进行的两个率比较检验的资料，都可用 χ^2 检验，两者是等价的，即 $u^2 = \chi^2$；相异点：(1) u 检验可进行单侧检验；(2) 满足四格表 u 检验的资料，计算两个率之差的可信区间，可从专业上判断两率之差有无实际意义；(3) χ^2 检验可用于 2×2 列联表资料有无关联的检验。

2．$R\times C$ 表 χ^2 检验的适用条件及当条件不满足时可以考虑的处理方法是什么？

答案：$R\times C$ 表 χ^2 检验的适用条件是理论频数不宜过小，否则有可能产生偏性。当条件不满足时有三种处理方法：①增大样本例数使理论频数变大；②删去理论数太小的行或列；③将理论数太小的行或列与性质相近的邻行或邻列合并，使重新计算的理论频数变大。但②、③法都可能会损失信息或损害样本的随机性，因此应慎用。

（四）计算题

1．为研究静脉曲张是否与肥胖有关，观察 122 对同胞兄弟，每对同胞兄弟中有一个属肥胖，另一个属正常体重，记录得静脉曲张发生情况见表 8-2，试分析之。

表 8-2　122 对同胞兄弟静脉曲张发生情况

正常体重	肥胖		合计
	发生	未发生	
发　生	19	5	24
未发生	12	86	98
合　计	31	91	122

[评析] 这是一个配对设计的资料，因此用配对 χ^2 检验公式计算。

H_0：肥胖者与正常体重者的静脉曲张发生情况无差别

H_1：肥胖者与正常体重者的静脉曲张发生情况不同

$\alpha = 0.05$

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} = \frac{(|5-12|-1)^2}{5+12} = 2.12，\ \nu = 1$$

$\chi^2 = 2.11 < \chi^2_{0.05,1}$，$P>0.05$，尚不能认为静脉曲张与肥胖有关。

2．某卫生防疫站在中小学观察三种矫正近视眼措施的效果，近期疗效数据见表 8-3。试对这三种措施的疗效作出评价。

表 8-3　三种措施的近期有效率比较

矫治方法	有效人数	无效人数	合计	有效率（%）
夏天无眼药水	51	84	135	37.78
新医疗法	6	26	32	18.75
眼保健操	5	13	18	27.78
合计	62	123	185	33.51

[评析]

H_0：三种措施有效率相等

H_1：三种措施有效率不相等或不全相等

$a = 0.05$

$$c^2 = n(\sum \frac{A^2}{n_r n_c} - 1) = 185 \times \left(\frac{51^2}{62 \times 135} + \frac{84^2}{123 \times 135} + \frac{6^2}{62 \times 32} + \frac{26^2}{123 \times 32} + \frac{5^2}{62 \times 18} + \frac{13^2}{123 \times 18} - 1 \right) = 4.498$$

，n =（2-1）（3-1）=2

查表得 0.25>P>0.10，按$a = 0.05$水准不拒绝H_0，尚不能认为三种措施有效率有差别。

3.某医院以 400 例自愿接受妇科门诊手术的未产妇为观察对象，将其分为 4 组，每组 100 例，分别给予不同的镇痛处理，观察的镇痛效果见表 8-4，问 4 种镇痛方法的效果有无差异？

表 8-4　4 种镇痛方法的效果比较

镇痛方法	例数	有效率（%）
颈麻	100	41
注药	100	94
置栓	100	89
对照	100	27

[评析] 为了应用c^2检验，首先应计算出有效和无效的实际频数，列出计算表，见表 8-5。

表 8-5　4 种镇痛方法的效果比较

镇痛方法	有效例数	无效例数	合计
颈麻	41	59	100
注药	94	6	100
置栓	89	11	100
对照	27	73	100
合计	251	149	400

H_0：4 种镇痛方法的效果相同

H_1：4 种镇痛方法的效果不全相同

$a = 0.05$

$$c^2 = n(\sum \frac{A^2}{n_r n_c} - 1) = 400 \times \left(\frac{41^2}{251 \times 100} + \frac{59^2}{149 \times 100} + \ldots + \frac{73^2}{149 \times 100} - 1 \right) = 146.175,$$

n =（4-1）（2-1）=3

查表得 P <0.05，按$a = 0.05$水准拒绝H_0，接受H_1，即 4 种镇痛方法的效果不全相同。

四、习　题

（十一）　单项选择题

1. 关于样本率 p 的分布正确的说法是：

A．服从正态分布

B．服从 c^2 分布

C．当 n 足够大，且 p 和 1-p 均不太小，p 的抽样分布逼近正态分布

D．服从 t 分布

2. 以下说法正确的是：

A．两样本率比较可用 u 检验

B．两样本率比较可用 t 检验

C．两样本率比较时，有 $u = c^2$

D．两样本率比较时，有 $t^2 = c^2$

3. 率的标准误的计算公式是：

A．$\sqrt{p(1-p)}$　　B．$\frac{p(1-p)}{n}$　　C．$\sqrt{\frac{p}{n-1}}$　　D．$\sqrt{\frac{p(1-p)}{n}}$

4. 以下关于 c^2 检验的自由度的说法，正确的是：

A．拟合优度检验时，$n = n-2$ （n 为观察频数的个数）

B．对一个 3×4 表进行检验时，$n = 11$

C．对四格表检验时，n =4

D．若 $c^2_{0.05,n} > c^2_{0.05,h}$，则 $n > h$

5. 用两种方法检查某疾病患者 120 名，甲法检出率为 60%，乙法检出率为 50%，甲、乙法一致的检出率为 35%，问两种方法何者为优？

A．不能确定　　B．甲、乙法一样　　C．甲法优于乙法　　D．乙法优于甲法

6．已知男性的钩虫感染率高于女性。今欲比较甲乙两乡居民的钩虫感染率，适当的方法是：

A．分性别比较　　B．两个率比较的 c^2 检验

C．不具可比性，不能比较　　D．对性别进行标准化后再做比较

7．以下说法正确的是

A．两个样本率的比较可用 u 检验也可用 c^2 检验

B．两个样本均数的比较可用 u 检验也可用 c^2 检验

C．对于多个率或构成比的比较，u 检验可以替代 c^2 检验

D．对于两个样本率的比较，c^2 检验比 u 检验可靠

（十二）名词解释

1．实际频数与理论频数

2．c^2 界值表

3．拟合优度

4．配对四格表

5．双向有序分类资料

6．率的标准误

7．多个率的两两比较

8．Fisher 精确概率

9． McNemar 检验

10． Yates 校正

（十三） 是非题

四个样本率做比较，$c^2 > c^2_{0.05(3)}$ ，可认为各总体率均不相等。

（十四） 计算题

1．121 名前列腺癌患者中，82 名接受电切术治疗，术后有合并症者 11 人；39 名接受开放手术治疗，术后有合并症 1 人。试分析两种手术的合并症发生率有无差异？

2．某厂在冠心病普查中研究冠心病与眼底动脉硬化的关系，资料见表 8-6。问两者是否存在一定的关系？

表 8-6 冠心病诊断结果与眼底动脉硬化级别的关系

眼底动脉硬化级别	冠心病诊断结果			合计
	正常	可疑	冠心病	
0	340	11	6	357
I	73	13	6	92
II	97	18	18	133
III	3	2	1	6
合计	513	44	31	588

3．表 8-7 是用两种方法检查已确诊的乳腺癌患者 120 名的检查结果，问：两种方法何者为优？

表 8-7 两种方法检查结果比较

乙法	甲法		合计
	+	-	
+	42	18	60
-	30	30	60
合计	72	48	120

4．用噬菌体治疗小儿细菌性痢疾结果见表 8-8，问两组阴转率有无显著差异？

表 8-8 两种方法检查结果比较

组 别	观察人数	粪见检阴性人数	阴转率（%）
试验组	29	25	86.2
对照组	28	17	60.7
合 计	57	42	73.7

5．某医院用冠心 2 号方治疗心绞痛患者，经三个月疗程后，疗效见表 8-9，问三个疗程组的有效率之间有无显著差异？

表 8-9 冠心 2 号方治疗心绞痛的有效率

疗 程	例数	有效例数	有效率（%）
一疗程	110	82	74.5
二疗程	150	130	86.7
三疗程	63	56	88.9
合 计	323	268	83.0

6．某医院比较急性黄疸型肝炎与正常人在超声波波型上的表现，见表 8-10。问两组肝波型的差异有无显著性？

表 8-10　急性黄疸型肝炎与正常人的超声波波形

组别	波　型			合计
	正常	可疑	较密	
黄疸型肝炎组	12	43	232	287
正 常 人 组	277	39	11	327
合计	289	82	243	614

7．有人研究惯用手与惯用眼之间是否存在一定关系，得资料如表 8-11，试作统计分析。

表 8-11　冠心 2 号方治疗心绞痛的有效率

	惯用左眼	两眼并用	惯用右眼	合计
惯用左手	34	62	28	124
两手并用	27	28	20	75
惯用右手	57	105	52	214
合　计	118	195	100	413

8.苏格兰西南部两个地区献血人员的血型记录如下表（表 8-12），问两地的血型分布是否相同？

表 8-12　两个地区献血人员的血型分布

地区	血　型				合计
	A	B	O	AB	
Eskdale	33	6	56	5	100
Annandale	54	14	52	5	125
合计	87	20	108	10	225

五、习题答题要点

（一） 单项选择题

1.C　2.A　3.D　4.D　5.A　6.D　7.A

（二） 名词解释

1．实际频数：actual frequency，即实际观察值。理论频数：theoretical frequency，在假设多个率或构成比相等的前提下，由合计率（构成比）推算出来的频数。

2．c^2 界值表：将 c^2 分布右侧尾部面积等于 a 时所对应的 c^2 值称为 c^2 分布的临界值，对于不同的自由度及 a 有不同的临界值，由这些临界值构成的表即 c^2 界值表。

3．拟合优度 ：goodness of fit，指一种度量某事物的频数分布是否符合某一理论分布或数据是否与模型吻合的方法。

4．配对四格表：为了控制随机误差而采用配对设计方案，将条件相似的两个受试对象配成一对，然后随机地让其中一个接受 A 处理，另一个接受 B 处理，每种处理的反应都按二项分类。全部 n 对实验结果的资料以表 8-12 表示，这样的表称为配对四格表。

表 8-12　配对四格表的形式

A 处理	B 处理	
	+	-
+	a	b
-	c	d

5. **双向有序分类资料：对于** $R\times C$ **表资料，当两个定性变量都有序时，这样的资料称为双向有序分类资料，如“急性放射病分度与放射烧伤面积占不同体表面积的百分比”，这里的两个变量均为有序的。**

6. **率的标准误：用以衡量由于抽样引起的样本率与总体率之间的误差的统计量，记为** s_P **。** $s_p=\sqrt{\frac{p(1-p)}{n}}$ **，** p **为总体率，** n **为样本容量；当总体率** p **未知时，以样本率** P **作为** p **的估计值，率的标准误为** $S_p=\sqrt{\frac{p(1-p)}{n}}$ **。**

7. **多个率的两两比较：指当假设检验确定了多个率之间存在差别后，检验哪两个两个样本率之间的差别具有统计学意义的方法。**

8. Fisher **精确概率：指当四格表中出现理论数小于** 1 **或** $n<40$ **时，用** R.A.Fisher（1934）**提出的方法直接计算出的有利于拒绝** H_0 **的概率。**

9. McNemar **检验：**McNemar's test for correlated proportions**，是分析配对四格表资料的方法，其计算公式为** $c^2=\frac{(|b-c|-1)^2}{b+c}$ **，** v=1**。**

10．Yates **校正：英国统计学家** Yates F **认为，由于** c^2 **分布理论上是一连续性分布，而分类资料是间断性的，由此计算出的** c^2 **值不连续，尤其是自由度为** 1 **的四格表，求出的概率** P **值可能偏小，此时需对** c^2 **值作连续性校正（**correction of continuity**），这一校正即所谓的** Yates **校正（**Yates' correction**）。**

（三） 是非题

错。多个样本率做比较时， H_1 **为各总体率不全相等，所以当接受** H_1 **时，并不能说明各总体率均不相等。**

（四） 计算题：

1．**将资料整理成四格表**

手术方法	合并症		
	+	-	
电切术	11	71	82
开放手术	1	38	39
	12	109	121

用四格表校正公式算得 $c^2=2.37$ **，** P>0.05, **尚不能认为两种手术的合并症发生率有差异。**

2．**该资料属双向有序分类资料,用** c^2 **检验解决。**

H_0: **冠心病诊断结果与眼底动脉硬化级别无关联**

H_1: **冠心病诊断结果与眼底动脉硬化级别有关联**

$a=0.05$

c^2=61.59，　$c^2<c^2_{0.01,6}$，P<0.05,

按$a=0.05$水准拒绝H_0接受H_1,故可认为冠心病与眼底动脉硬化有关联。

3．采用配对c^2检验。

H_0: 两法不分优劣

H_1: 两法能分优劣　$a=0.05$

c^2=3.00，按$a=0.05$水准不拒绝H_0，尚不能认为检出率有差别。

4．可用 u 检验或c^2检验。用 c^2检验时，首先将资料整理成四格表形式，然后再代入公式。算得c^2=4.774，按$a=0.05$水准拒绝H_0接受H_1,认为两组阴转率差别有统计学意义。

5．用$R\times C$表c^2检验公式算得c^2=8.539，v=2，P<0.05，，按$a=0.05$水准拒绝H_0接受H_1，三个疗程有效率的差异有统计学意义。

6．用$R\times C$表c^2检验公式算得c^2=443.456，v=2，P<0.05，，按$a=0.05$水准拒绝H_0接受H_1，两组肝波型的差异有统计学意义。

7．由c^2检验公式算得c^2=4.020，v=4，P>0.05，，按$a=0.05$水准不拒绝H_0，尚不能认为惯用手与惯用眼之间存在关系。

8．本例只有一个格子的理论频数小于 5，故仍可用c^2检验。c^2=5.710，v=3，P>0.05，，按$a=0.05$水准不拒绝H_0，尚不能认为两地的血型分布不同。

（徐勇勇　马跃渊）

第九章 秩和检验

一、教学大纲要求

（一）掌握内容

1．非参数统计基本概念和特点。

2．配对设计差值的符号秩检验。

3．成组设计资料两样本比较的秩和检验。

（二）熟悉内容

1．成组设计多样本比较的秩和检验步骤。

2．随机区组设计资料的秩和检验。

（三）了解内容

1．成组设计多样本两两比较的秩和检验。

2．随机区组设计资料两两比较的秩和检验。

二、教学内容精要

（一）参数统计与非参数统计

1．**参数统计**

样本所来自的总体分布具有某个已知的函数形式，而其中有的参数是未知的，统计分析的目的就是对这些未知的参数进行估计或检验。此类方法称为参数统计。

2．**非参数统计**

样本所来自的总体分布难以用某种函数式来表达，还有一些资料的总体分布的函数式是未知的，只知道总体分布是连续型的或离散型的，解决这类问题的一种不依赖总体分布的具体形式的统计方法。由于这类方法不受总体参数的限制，故称非参数统计法（non-parametric statistics**），或称为不拘分布（**distribution-free statistics**）的统计分析方法，又称为无分布型式假定（**assumption free statistics**）的统计分析方法。它检验的是分布，而不是参数。非参数统计不需对总体分布(总体参数)作出特殊假设。**

（二）非参数统计的特点和适用范围

1．**特点**

（1）样本所来自的总体的分布形式为任何形式，甚至是未知的，都能适用。

（2）收集资料方便，可用“等级”或“符号”来评定观察结果。

（3）多数非参数方法比较简便，易于理解和掌握。

（4）缺点是损失信息量，适用于参数统计法的资料用非参数统计方法进行检验将降低检验效能。

2．适用范围

（1）等级资料。

（2）偏态分布资料。当观察资料呈偏态或极度偏态分布而又未作变量变换，或虽经变量变换仍未达到正态或近似正态分布时，宜用非参数检验。

（3）各组离散程度相差悬殊，即方差明显不齐，且不能变换达到齐性。

（4）个别数据偏离过大，或资料为单侧或双侧没有上限或下限值。

（5）分布类型不明。

（6）初步分析。有些医学资料由于统计工作量大，可采用非参数统计方法进行初步分析，挑选其中有意义者再进一步分析(包括参数统计内容)。

（7）对于一些特殊情况，如从几个总体所获得的数据，往往难以对其原有总体分布作出估计，在这种情况下可用非参数统计方法。

（三）配对设计差值的符号秩检验(Wilcoxon配对法)

1.检验步骤

(1)假设：H_0：差值总体中位数M_d=0

H_1：$M_d \neq 0$

$\alpha = 0.05$

(2)求差值

(3)编秩：依差值的绝对值从小到大编秩。编秩时遇差数等于 0，舍去不计，同时样本例数减1；遇绝对值相等差数，符号相同顺次编秩，符号相反取平均秩次，且符号相反。

(4)求秩和并确定检验统计量：分别求出正负秩次之和，正秩和以 T_+表示，负秩和的绝对值以 T-表示。T_+及 T-之和应等于 $n(n+1)/2$，任取 T_+(或 T-)作检验统计量 T 。

(5)确定 P 值和作出推断结论：当 $n \leq 50$ 时，查 T 界值表，得出 P 值。若检验统计量 T 值在上、下界值范围内，其 P 值大于表上方相应概率水平；若 T 值在上、下界值上若范围外，其 P 值小于表上方相应概率水平。

2．正态近似法

若 $n>50$ 时，可用 u 检验，按如下公式计算 u 值：

$$u = \frac{|T - n(n+1)/4| - 0.5}{\sqrt{n(n+1)(2n+1)/24}} \qquad (9\text{-}1)$$

当相同差值数多时，应改用校正式：

$$u = \frac{|T - n(n+1)/4| - 0.5}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum(t_j^3 - t_j)}{48}}} \qquad (9\text{-}2)$$

（四）成组设计两样本比较的秩和检验(Wilcoxon两样本比较法)

1.检验步骤：

(1)假设：H_0：两总体分布相同

H_1：两总体分布不同

$\alpha = 0.05$

(2)编秩：将两组原始数据分别由小到大排队，再将原始数据从小到大统一编秩。编秩时遇同组相同数据，顺次编秩，遇不同组相同数据取平均秩次。

(3)求秩和并确定检验统计量：当两样本例数不等时，以样本例数小者为 n_1，其秩和为 T。相等时，可任取一组的秩和为 T。

(4)确定 P 值和作出推断结论：查 T 界值表，得出 P 值。若检验统计量 T 值在上、下界

值范围内，其 P 值大于表上方相应概率水平；若 T 值在上、下界值上若范围外，其 P 值小于表上方相应概率水平。

2.正态近似法

若 n_1 或 n_2-n_1 较大时，可用 u 检验，按如下公式计算 u 值：

$$u=\frac{|T-n_1(N+1)/2|-0.5}{\sqrt{n_1n_2(N+1)/12}} \tag{9-3}$$

当相同差值数多时，应改用校正式：

$$u_c=u\Big/\sqrt{C} \tag{9-4}$$

其中：$C=1-\sum(t_j^3-t_j)\Big/(N^3-N)$　t_j 为第 j 个相同秩次的个数。

（五）成组设计多个样本比较的秩和检验(Kruskal-Wallis 法)

检验步骤：

1．假设：H_0：各总体分布相同

H_1：各总体分布不同

$a=0.05$

2．编秩：将两组原始数据分别由小到大排队，再将原始数据从小到大统一编秩。编秩时遇同组相同数据，顺次编秩，遇不同组相同数据取平均秩次。

3．求秩和并确定检验统计量：将各组秩次相加。

4．计算检验统计量 H 值：

$$H=\frac{12}{N(N+1)}(\sum\frac{R_i^2}{n_i})-3(N+1) \tag{9-5}$$

若各样本相同秩次较多时，应用校正公式H_c：

$$H_c=H/C \tag{9-6}$$

其中：$C=1-\sum(t_j^3-t_j)\Big/(N^3-N)$　t_j 为第 j 个相同秩次的个数。

5．确定 P 值和作出推断结论：查 H 界值表，得出 P 值。若检验统计量 T 值在上、下界值范围内，其 P 值大于表上方相应概率水平；若 T 值在上、下界值上若范围外，其 P 值小于表上方相应概率水平。

（六）多个样本两两比较的秩和检验(Nemenyi 法)

检验步骤：

1．假设：H_0：各总体分布相同

H_1：任意两总体的位置不同

$a=0.05$

2．求秩和的差值：计算各组中所有可能两两对比组秩和差数的绝对值$D=|R_A-R_B|$

3．确定 P 值和作出推断结论：（1）当各样本例数相等时，查 D 界值表或计算界值，得出 P 值。（2）当各样本例数不等或不全等时，将各对比组平均秩次之差与界值比较，界值计

算公式如下：

$$c^2 = \frac{|\overline{R}_A - \overline{R}_B|}{C[N(N+1)/12][1/n_A + 1/n_B]} \qquad (9\text{-}7)$$

其中：相同秩次校正数 $C = 1 - \sum(t_j^3 - t_j)/(N^3 - N)$ **t_j 为第 j 个相同秩次的个数；** $c^2_{a,(k-1)}$ **查 χ^2 界值表；N 为各处理组的总例数。**

（七）随机区组设计资料的秩和检验

1.**查表法**

检验步骤：

(1)**将每个区组的数据由小到大分别编秩，遇相同数值取平均秩；**

(2)**计算各处理组的秩和** R_i **；**

(3)**求平均秩**:$R=b(k+1)/2$　**式中，**b**：区组数**　k**：处理组数；**

(4)**计算各处理组的**(R_i-R);

(5)**求** $M=\Sigma(R_i-R)^2$

(6)**查** M **界值表，** M **大于或等于表中数值则差别有统计意义。**

2. Friedman**检验**

检验步骤：

(1)**将各区组内数据由小到大分别编秩，遇相同数值取平均秩次**

(2)**计算各处理组的秩和** R_i **；**

若各区组内无相同秩次，可用：

$$c^2 = \frac{12}{bk(k+1)}\sum_{j=1}^{k} R_j^2 - 3b(k+1) \qquad (9\text{-}8)$$

(3)**查** $c^2_{a,(k-1)}$ **界值，确定** P **值，作出推断。**

（八）随机区组设计资料的两两比较

检验步骤：

(1)**计算各处理组的秩和** R_i **；**

(2)**计算各对比组秩和的差：** $\overline{R}_A - \overline{R}_B$

$$u = \frac{|\overline{R}_A - \overline{R}_B|}{\sqrt{bk(k+1)/6}} \qquad (9\text{-}9)$$

查 u **界值，确定** P **值，若共进行** c **次比较，则用** α/c **作检验水平，作出推断。**

三、典型试题分析

（一）单项选择题

1．以下对非参数检验的描述哪一项是错误的是（　）。

A. **非参数检验方法不依赖于总体的分布类型**

B. **应用非参数检验时不考虑被研究对象的分布类型**

C. **非参数的检验效能低于参数检验**

D. **一般情况下非参数检验犯第二类错误的概率小于参数检验**

答案：D

[评析]　本题考点：非参数检验的特点。

非参数检验优点是应用范围广、简便、易掌握、不依赖于总体分布；缺点是若资料符合参数检验条件而用非参数检验，则检验效率低于参数检验。

2．多样本计量资料比较，当分布类型不清时选择（　）。

A. t **检验**　　B. u **检验**

C. **秩和检验**　　D. χ^2 **检验**

答案：C

[评析]　本题考点：非参数检验的适用范围。

分布类型不明时，差别检验应首先考虑非参数统计方法。

3．符合 t **检验条件的数值变量资料如果采用秩和检验，不拒绝** H_0 **时（　）。**

A．**第一类错误增大**　　B．**第二类错误增大**

C．**第一类错误减少**　　D．**第二类错误减少**

答案：B

[评析]　本题考点：非参数检验与非参数检验的区别。

当资料符合参数检验条件时，非参数检验检验效能要比参数检验低，发现总体差异的能力不如参数检验高，容易把一些本来有差别的总体检验成同一总体。

4．按等级分组的资料作秩和检验时，如果用 H **值而不用校正后的** H_c **值，则会（　）。**

A．**提高检验的灵敏度**

B．**会把一些无差别的总体推断成有差别**

C．**会把一些有差别的总体推断成无差别**

D．**第一、二类错误概率不变**

答案：C

[评析] 本题考点：Kruskal-wallis 秩和检验校正公式的应用。

当各样本相同秩次较多时，应用校正公式 Hc：$H_c = H / C$ 其中：$C = 1 - \sum(t_j^3 - t_j) \big/ (N^3 - N)$

t_j 为第 j 个相同秩次的个数。由于 C<1，因此 H_C>H，所求得相应概率 P 要大一些，那么就会把一些有差别的总体推断成无差别。

(二)简答题

“对某资料进行统计分析时，应尽量采用参数检验方法，一般不易采用非参数检验方法”，试评价这种说法正确否？

答案：应根据设计的方案、资料性质和分析过程中所遇到的实际情况等来确定采用何种统计检验方法。当资料满足参数检验方法时，必须使用参数检验方法。反之，当资料不满足参数检验方法时，如资料分布不明、呈偏态分布、方差不齐、等级资料时，必须采用非参数检验方法。在实际工作中，许多资料不满足参数检验的条件，非参数检验并不比参数检验应用的场合少。所以，以上说法不正确。

四、习 题

（二） 名词解释

1．非参数统计　　2．参数统计　　3．秩次　　4．秩和

（二） 单项选择题

1. 以下检验方法之中，不属于非参数检验法的是（ ）。

A． t 检验　　B． 符号检验

C． Kruskal-Wallis 检验　　D． Wilcoxon 检验

2. 以下对非参数检验的描述哪一项是错误的（ ）。

A．参数检验方法不依赖于总体的分布类型

B．应用非参数检验时不考虑被研究对象的分布类型

C．非参数的检验效能低于参数检验

D．一般情况下非参数检验犯第二类错误的概率小于参数检验

3. 符合方差分析检验条件的成组设计资料如果采用秩和检验，则（ ）。

A．一类错误增大　　B．第二类错误增大

C．第一类错误减小　　D．第二类错误减小

4. 等级资料的比较宜用（ ）。

A．t 检验　　B．秩和检验

C．F检验　　D．四格表 X^2 检验

5. 在进行成组设计两样本秩和检验时，以下检验假设正确的是（ ）。

A．H_0：两样本对应的总体均数相同

B．H_0：两样本均数相同

C．H_0：两样本对应的总体分布相同

D．H_0：两样本的中位数相同

6． 在进行Wilcoxon配对法秩和检验时，以下检验假设正确的是（ ）。

A．H_0：两样本对应的总体均数相同

B．H_0：两样本的中位数相同

C．H_0：两样本对应的总体分布相同

D．以上都不正确

7． 两个小样本比较的假设检验，应首先考虑（ ）。

A．t检验　　B．秩和检验

C．任选一种检验方法　　D．资料符合哪种检验的条件

8． 对于配对比较的秩和检验，其检验假设为（ ）。

A．样本的差数应来自均数为0的正态总体

B．样本的差数应来自均数为0的非正态总体

C．样本的差数来自中位数为0的总体

D．样本的差数来自方差齐性和正态分布的总体

9． 在配对比较的差数秩和检验中，如果有两个差数为0，则（ ）。

A．对正秩和有0.5和1，对负秩和有-0.5和-1

B．对正秩和有2，对负秩和有-2

C．对正秩和有3，对负秩和有-3

D．不予考虑

10． 在成组资料的秩和检验中，设样本为n个秩号：1，2，...，n，如果有相同秩号，比如：i 和 i+1变成两个i+0.5，则对样本所有秩号的均数和标准差的影响有（ ）。

A．均数不变，标准差不变　　B．均数不变，标准差加大

C．均数加大，标准差减小　　D．均数不变，标准差减小

11． 若随机化成组设计资料来自于正态总体，分别采用秩和检验与t检验、u检验，则它们检验效率关系正确的是（ ）。

A．t检验>u检验>秩和检验　　B．u检验>秩和检验>t检验

C．t检验>秩和检验>u检验　　D．t检验，u检验>秩和检验

12． 配对比较的秩和检验的基本思想是：如果检验假设成立，则对样本来说（ ）。

A．正秩和的绝对值小于负秩和的绝对值

B．正秩和的绝对值大于负秩和的绝对值

C．正秩和的绝对值与负秩和的绝对值不会相差很大

D．正秩和的绝对值与负秩和的绝对值相等

13． 按等级分组资料的秩和检验中，各等级平均秩次为（ ）。

A．该等级的秩次范围的上界

B．该等级的秩次范围的下界

C．该等级的秩次范围的上界、下界的均数

D．该等级的秩次范围的上界、下界的之和

14． 成组设计多组资料比较的秩和检验，确定 P 值时，可利用查表法的情况正确的是（ ）。

A．组数<3,每组例数<5　　B．组数<3,每组例数≤5

C．组数≤3,每组例数<5　　D．组数≤3,每组例数≤5

15．配对设计资料的秩和检验,确定 P 值时,可利用查表法的样本例数 n 的范围为(　)。

A．50≥n≥5　　B．30≥n≥5

C．30≥n≥3　　D．50≥n≥3

16．成组设计两样本资料的秩和检验，样本例数分别为 n_1、n_2 ,按检验水准为 0.05（双侧)，可利用查表法确定显著性水平的情况正确的是（　)。

A．n_1=4　，n_2=4　　B．n_1=2　，n_2=4

C．n_1=9　，n_2=20　　D．n_1=11 ，n_2=11

17．非参数统计应用条件是（　)。

A．总体是正态分布

B．若两组比较，要求两组的总体方差相等

C．不依赖于总体分布

D．要求样本例数很大

18．下述哪些不是非参数统计的特点（　)。

A．不受总体分布的限定　　B．多数非参数统计方法简单，易于掌握

C．适用于等级资料　　D．检验效能总是低于参数检验

19．设配对设计资料的变量值为 X_1 和 X_2，则配对资料的秩和检验（　)。

A．把 X_1 与 X_2 的差数绝对值从小到大编秩

B．把 X_1 和 X_2 综合从小到大编秩

C．把 X_1 和 X_2 综合按绝对值从小到大编秩

D．把 X_1 与 X_2 的差数从小到大编秩

20．秩和检验和 t 检验相比，其优点是（　)。

A．计算简便，不受分布限制　　B．公式更为合理

C．检验效能高　　D．抽样误差小

21．配对设计差值的符号秩检验，对差值编秩时，遇有差值绝对值相等时（　)。

A．符号相同，则取平均秩次　　B．符号相同，仍按顺序编秩

C．符号不同，仍按顺序编秩　　D．不考虑符号，按顺序编秩

22．配对设计的秩和检验中，其 H_0 假设为（　)。

A．差值的总体均数为 0　　B．差值的总体中位数为 0

C．μ_d?0　　D．M_d?0

23．一组 n_1 和一组 n_2（n_2>n_1）的两个样本资料比较，用秩和检验，有（　)。

A．n_1 个秩次 1，2，...，n_1

B．n_2 个秩次 1，2，...，n_2

C．n_1+n_2 个秩次 1，2，...，n_1+n_2

D．n_1-n_2 个秩次 1，2，...，n_1-n_2

24．成组设计两样本比较的秩和检验中，描述不正确的是（　)。

A．将两组数据统一由小到大编秩

B．遇有相同数据，若在同一组，按顺序编秩

C．遇有相同数据，若不在同一组，按顺序编秩

D．遇有相同数据，若不在同一组，取其平均秩次

25．成组设计的两小样本均数比较的假设检验（　）。

A．t 检验

B．成组设计两样本比较的秩和检验

C．t 检验或成组设计两样本比较的秩和检验

D．资料符合 t 检验条件还是成组设计两样本比较的秩和检验条件

26．对两样本均数作比较时，已知 n_1、n_2 均小于 30，总体方差不齐且分布呈偏态，宜用（　）。

A．t 检验　　B．u 检验

C．秩和检验　　D．F 检验

27．等级资料两样本比较的秩和检验中，如相同秩次过多，应计算校正 u_c 值，校正的结果使（　）。

A．u 值增加，P 值减小　　B．u 值增加，P 值增加

C．u 值减小，P 值增加　　D．u 值减小，P 值减小

28．符号秩检验（Wilcoxon 配对法）中，秩和 T 和 P 值的关系描述正确的是（　）。

A．T 落在界值范围内，则 P 值大于相应概率

B．T 落在界值范围上界外，则 P 值大于相应概率

C．T 落在界值范围下界外，则 P 值大于相应概率

D．T 落在界值范围上，则 P 值大于相应概率

29．配对设计资料的符号秩检验中，如相同秩次过多，未计算校正 u_c 值，而计算 u 值，不拒绝 H_0 时（　）。

A．第一类错误增加　　B．第一类错误减少

C．第二类错误增加　　D．第二类错误减小

（三）是非题

1．统计资料符合参数检验应用条件，但数据量很大，可以采用非参数方法进行初步分析。

2．对同一资料和同一研究目的，应用参数检验方法，所得出的结论更为可靠。

3．等级资料差别的假设检验只能采用秩和检验，而不能采用列联表 χ^2 检验等检验方法。

4．非参数统计方法是用于检验总体中位数、极差等总体参数的方法。

（四）计算题

1．下表资料是 8 名健康成年男子服用肠溶醋酸棉酚片前后的精液检查结果，服用时间为 1～3 个月，问服药后精液中精子浓度有无下降？

表 9-1　服药前后精子浓度(万/ml)

编号	1	2	3	4	5	6	7	8
服药前	6000	22000	5900	4400	6000	6500	26000	5800
服药后	660	5600	3700	5000	6300	1200	1800	2200

2．某营养实验室随机抽取 24 只小鼠随机分为两组，一组饲用未强化玉米，一组饲用已

强化玉米，观察玉米强化前后干物质可消化系数的差别有无显著意义。

表 9-2　玉米干物质可消化系数

已强化组		未强化组	
可消化系数（%）	秩次	可消化系数	秩次
34.3		<10	
38.1		15.8	
42.8		18.2	
45.9		21.9	
48.2		23.4	
51.7		24.6	
52.4		26.1	
52.8		27.2	
54.5		29.3	
54.8		30.7	
55.3		34.4	
65.4		34.7	
秩和	T_1=		T_2=

3．配对设计的两组鼠肝中维生素 A 含量（IU/g）有无显著差异，用秩和检验和 *t* 检验分别作检验，试比较两法的检验结果并加以说明。

表 9-3 不同饲料组鼠肝维生素 A 含量

大鼠配偶组	肝中维生素 A 含量		差数
	正常饲料组	维生素 E 缺乏组	*d*
1	3550	2450	1100
2	2000	2400	-400
3	3000	1800	1200
4	3950	3200	750
5	3800	3950	-150
6	3750	2700	1050
7	3450	2500	950
8	3050	1750	1300
9	2500	2550	-50
10	3650	3750	-100

4．以下是测得的铅作业与非铅作业工人的血铅值(μmol/L)，请问两组工人的血铅值有无差别？

表 9-4 两组工人血铅测定值(μmol/L)

患者	0.82	0.87	0.97	1.21	1.64	2.08	2.13			
健康人	0.24	0.24	0.29	0.33	0.44	0.58	0.63	0.72	0.87	1.01

5．在研究人参镇静作用的实验中，曾有人以 5%人参浸液对某批小白鼠 20 只作腹腔注射，而以等量蒸馏水对同批 12 只小白鼠作同样注射为对照，问能否说人参有显著的镇静作用？

表 9-5 人参镇静作用的实验结果

镇静等级	例数	
	人参组	对照组
-	4	11
±	1	...
+	2	1
++	1	...
+++	12	...

五、习题答题要点

(一)名词解释

1．非参数统计：针对某些资料的总体分布难以用某种函数式来表达，或者资料的总体分布的函数式是未知的，只知道总体分布是连续型的或离散型的，用于解决这类问题的一种不依赖总体分布的具体形式的统计分析方法。由于这类方法不受总体参数的限制，故称非参数统计法(non-parametric statistics)，或称为不拘分布(distribution-free statistics)的统计分析方法，又称为无分布型式假定（assumption free statistics）的统计分析方法。

2．参数统计：通常要求样本来自总体分布型是已知的（如正态分布），在这种假设的基础上，对总体参数（如总体均数）进行估计和检验，称为参数统计(parametric statistics)

3．秩次：变量值按照从小到大顺序所编的秩序号称为秩次（rank)。

4．秩和：各组秩次的合计称为秩和（rank sum)，是非参数检验的基本统计量。

（二）单项选择题

1.A 2.D 3.B 4.B 5.C 6.D 7.D 8.C 9.D 10.D

11.D 12.C 13.C 14.D 15.A 16.A 17.C 18.D 19.A 20.A

21.B 22.B 23.C 24.C 25.D 26.C 27.A 28.A 29.C

(三)是非题

1． 正确。

2．错误。应视资料的特性而定，若资料符合参数检验方法的条件，就运用参数检验方法；若符合非参数检验方法的条件，就运用非参数检验方法。

3．错误。应根据研究目的和资料性质而定，例如当资料的实验分组变量有序，而指标

分组变量无序时，可以采用列联表 χ^2 检验。

4．错误。非参数检验是检验总体分布，而非总体参数。

（四）计算题

1．答案：由于本资料数据离散程度相当大，分布不明，故宜用配对设计差值的符号秩检验（Wilcoxon配对法）。负秩和 T_-=4.5，正秩和 T_+=61.5，P<0.05。

2．答案：由于本资料中存在截尾数据，故宜用成组设计两样本比较的秩和检验（Wilcoxon两样本比较法）。第一组 n_1=12,秩和 T_1=220，第二组 n_2=12,秩和 T_2=80,P<0.01。

3．答案：本资料应用配对设计差值的符号秩检验（Wilcoxon 配对法）。负秩和 T_-=10，正秩和 T_+=45，P>0.05。若使用配对设计的 t 检验，则 t=2.711，P<0.05。由此可见，按检验水准为0.05时，二者检验结果不一致，此时，应对样本作正态性检验，若样本所来自的总体服从正态分布，则 t 检验结果更可取，否则，秩和检验的结果更加可靠。在本例中，经检验样本所来自的总体服从正态分布，故可以说不同饲料组鼠肝维生素 A 含量不同。

4．答案：由于本资料为成组设计，两组血铅方差不齐，故宜用成组设计两样本比较的秩和检验(Wilcoxon两样本比较法)。第一组 n_1=7,秩和 T_1=93.5，第二组 n_2=10,秩和 T_2=59.5，0.01<P<0.05。

5．答案：本资料应用成组设计两样本比较的秩和检验（Wilcoxon 两样本比较正态近似法）。第一组 n_1=20,秩和 T_1=422，第二组 n_2=12,秩和 T_2=106，u_c=3.76， P<0.01。

（夏结来　蒋红卫）

第十章 直线相关与回归

一、教学大纲要求

（一） 掌握内容

1. 直线相关与回归的基本概念。

2. 相关系数与回归系数的意义及计算。

3. 相关系数与回归系数相互的区别与联系。

（二）熟悉内容

1. 相关系数与回归系数的假设检验。

2. 直线回归方程的应用。

3. 秩相关与秩回归的意义。

（三）了解内容

曲线直线化。

二、 学内容精要

(一) 直线回归

1. **基本概念**

直线回归(linear regression)**建立一个描述应变量依自变量变化而变化的直线方程，并要求各点与该直线纵向距离的平方和为最小。直线回归是回归分析中最基本、最简单的一种，故又称简单回归（**simple regression**）。**

直线回归方程$\hat{Y}=a+bX$ **中，**a**、**b**是决定直线的两个系数，见表** 10-1**。**

表 10-1 **直线回归方程** a**、**b**两系数对比**

	a	b
含义	**回归直线在** Y**轴上的截距**（intercept）**。表示** X**为零时，** Y**的平均水平的估计值。**	**回归系数**（regression coefficient）**，即直线的斜率。表示** X**每变化一个单位时，** Y**的平均变化量的估计值。**
系数>0	a>0 **表示直线与纵轴的交点在原点的上方**	b>0**，表示直线从左下方走向右上方，即** Y**随** X**增大而增大**
系数<0	a<0 **表示直线与纵轴的交点在原点的下方**	b<0**，表示直线从左上方走向右下方，即** Y**随** X**增大而减小**
系数=0	a=0 **表示回归直线通过原点**	b=0**，表示直线与** X**轴平行，即** Y**不随** X**的变化而变化**
计算公式	$a=\overline{Y}-b\overline{X}$	$b=\dfrac{\sum(X-\overline{X})(Y-\overline{Y})}{\sum(X-\overline{X})^2}=\dfrac{l_{XY}}{l_{XX}}$

2. 样本回归系数 b 的假设检验

（1）方差分析；

（2）t 检验。

3. 直线回归方程的应用

（1）描述两变量的依存关系；

（2）用回归方程进行预测；

（3）用回归方程进行统计控制；

（4）用直线回归应注意的问题。

(二) 直线相关

1. 基本概念

直线相关（linear correlation）又称简单相关（simple correlation），用于双变量正态分布资料。有正相关、负相关和零相关等关系。直线相关的性质可由散点图直观的说明。

相关系数又称积差相关系数（coefficient of product-moment correlation），以符号 r 表示样本相关系数，ρ 表示总体相关系数。它是说明具有直线关系的两个变量间，相关关系的密切程度与相关方向的指标。

2. 计算公式

$$r=\frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}=\frac{l_{XY}}{\sqrt{l_{XX}}\sqrt{l_{YY}}}$$

相关系数 r 没有单位，其值为 $-1\le r\le 1$。其绝对值愈接近 1，两个变量间的直线相关愈密切；愈接近 0，相关愈不密切。r 值为正表示正相关，说明一变量随另一变量增减而增减，方向相同；r 值为负表示负相关，说明一变量增加、另一变量减少，即方向相反；r 的绝对值等于 1 为完全相关。

3. 样本相关系数 r 的假设检验

（1）r 界值表法；

（2）t 检验法。

（三）直线回归与相关的区别与联系

1. 区别

（1） 资料要求：直线回归要求因变量 Y 服从正态分布，X 是可以精确测量和严格控制的变量，一般称为 I 型回归；直线相关要求两个变量 X、Y 服从双变量正态分布。这种资料若进行回归分析称为II型回归。

（2） 应用情况：直线回归是说明两变量依存变化的数量关系；直线相关是说明两变量间的相关关系。

（3） 意义：b 表示 X 每增（减）一个单位时，Y 平均改变 b 个单位；r 说明具有直线关系的两个变量间关系的密切程度与相关方向。

（4） 计算：$b=l_{xy}/l_{xx}$ ；$r=l_{xy}/\sqrt{l_{xx}l_{yy}}$ 。

（5） 取值范围：$-\infty<b<+\infty$ ；$-1\le r\le 1$ 。

（6） 单位：b 有单位；r 没有单位。

2. 联系

（1） 方向一致：对一组数据若能同时计算 b 和 r,它们的符号一致。

（2） 假设检验等价：对同一样本，r 和 b 的假设检验得到的 t 值相等，即 $t_b=t_r$ 。

（3） 用回归解释相关：决定系数 $r^2=\dfrac{l_{xy}^2}{l_{xx}l_{yy}}=SS_{回}/SS_{总}$ ，回归平方和越接近总平方和，则 r^2 越接近 1，说明引入相关的效果越好。

（四）秩相关

秩相关，又称等级相关（rank correlation），是用双变量等级数据作直线相关分析，适用于下列资料：

1. 不服从双变量正态分布而不宜作积差相关分析；
2. 总体分布型未知；
3. 用等级表示的原始数据。

三、典型试题分析

1．回归系数的假设检验（ ）

A．只能用 r 的检验代替　　B．只能用 t 检验

C．只能用 F 检验　　D．三者均可

答案：D

[评析] 本题考点：回归系数假设检验方法的理解。

回归系数的假设检验常用的方法有：①方差分析；②t 检验。对同一样本，r 和 b 的假设检验等价，r 和 b 的假设检验得到的 t 值相等，即 $t_b=t_r$。故回归系数的假设检验用三者均可。

2．已知 $r_1=r_2$，那么（ ）

A．$b_1=b_2$　　B．$t_{b1}=t_{b2}$

C．$t_{r1}=t_{r2}$　　D．两样本决定系数相等

答案： D

[评析] 本题考点：直线相关系数与回归系数关系的理解。

因为相关系数 r 和回归系数 b 的计算公式不同，不能推导出 $b_1=b_2$ ；r 和 b 的假设检验等价，即 $t_{r1}=t_{b1}$，$t_{r2}=t_{b2}$，而不是 $t_{b1}=t_{b2}$，$t_{r1}=t_{r2}$ ；样本决定系数为 r^2，已知 $r_1=r_2$，则两样本决定系数相等，即 $r_1^2=r_2^2$。

3．$|r|>r_{0.05(n-2)}$时，可认为两变量 X 与 Y 间（ ）

A．有一定关系　　B. 有正相关关系

C．一定有直线关系　　D. 有直线关系

答案： D

[评析] 本题考点：直线相关系数假设检验的理解。

因为直线相关系数 r 是样本的相关系数，它是相应总体相关系数 ρ 的估计值。由于抽样误差的影响，必须进行显著性检验。r 的假设检验是检验两变量是否有直线相关关系。$|r|>r_{0.05(n-2)}$时，$P<0.05$，拒绝 H_0，接受 H_1，认为总体相关系数 $\rho\neq0$，因此可认为两变量 X 与 Y 间有直线关系。

4．相关系数检验的无效假设 H_0 是（ ）

A．$\rho=0$　　B. $\rho\neq0$

C．$\rho>0$　　D. $\rho<0$

答案： A

[评析] 本题考点：直线相关系数显著性检验中检验假设的理解。

因为 r 是样本相关系数，它是总体相关系数 ρ 的估计值。要判两变量间是否有相关关系，就要检验 r 是否来自总体相关系数 ρ 为零的总体。因为即使从 $\rho=0$ 的总体作随机抽样，由于抽样误差的影响，所得 r 值也常不等于零。

5．同一双变量资料，进行直线相关与回归分析，有（　）。

A．$r>0$，$b<0$　　B. $r>0$，$b>0$

C．$r<0$，$b>0$　　D. r 与 b 的符号毫无关系

答案： B

[评析] 本题考点：直线相关与回归的区别与联系的理解。

因为对同一资料而言直线相关系数与回归系数的方向一致，若能同时计算 b 和 r,它们的符号一致。因此，同一双变量资料，进行直线相关与回归分析，有 $r>0$，$b>0$。

四、习　题

（三）单项选择题

19．　下列(　　)式可出现负值。

A．$\Sigma(X-\bar{X})^2$　　B．$\Sigma Y^2-(\Sigma Y)^2/n$

C．$\Sigma(Y-\bar{Y})^2$　　D．$\Sigma(X-\bar{X})(Y-\bar{Y})$

20．　$Y=14+4X$ 是 1~7 岁儿童以年龄（岁）估计体重（市斤）的回归方程，若体重换成国际单位 kg，则此方程(　　)。

A．截距改变　　B．回归系数改变

C．两者都改变　　D．两者都不改变

21．　已知 $r=1$，则一定有(　　)。

A．$b=1$　　B．$a=1$

C．$S_{Y.X}=0$　　D．$S_{Y.X}=S_Y$

22．　用最小二乘法确定直线回归方程的原则是各观察点(　　)。

A．距直线的纵向距离相等

B．距直线的纵向距离的平方和最小

C．与直线的垂直距离相等

D．与直线的垂直距离的平方和最小

23．　直线回归分析中，X 的影响被扣除后，Y 方面的变异可用指标(　　)表示。

A．$S_{x,y}=\sqrt{\Sigma(X-\bar{X})^2/(n-2)}$　　B. $S_r=\sqrt{\Sigma(Y-\hat{Y})^2/(n-1)}$

C. $S_{y,x}=\sqrt{\Sigma(Y-\hat{Y})^2/(n-2)}$　　D. $S_b=S_{xy}\Big/\sqrt{\Sigma\left(X-\bar{X}\right)^2}$

24．　直线回归系数假设检验，其自由度为(　　)。

A．n　　B．$n-1$

C．$n-2$　　D．$2n-1$

25． 应变量 Y 的离均差平方和划分，可出现(　　)。

A．$SS_{剩}=SS_{回}$　　B．$SS_{总}=SS_{剩}$

C．$SS_{总}=SS_{回}$　　D．以上均可

26． 下列计算 $SS_{剩}$ 的公式不正确的是(　　)。

A．$l_{YY}-l_{XY}b$　　B． $l_{YY}-bl_{XX}$

C. $l_{YY}-l_{XY}^2/l_{XX}$　　D．$(1-r^2)l_{YY}$

27． 直线相关系数可用(　　)计算。

A．$l_{XY}/\sqrt{l_{XX}l_{YY}}$　　B．$b_{YX}\sqrt{l_{XX}/l_{YY}}$

C．$\sqrt{b_{YX}b_{XY}}$　　D． 以上均可

28． 当 $r=0$ 时，$\hat{Y}=a+bX$ 回归方程中有(　　)。

A．a 必大于零　　B. a 必等于 $\bar{X}$

C．a 必等于零　　D. a 必等于 $\bar{Y}$

（四）名词解释

1．直线回归　2．回归系数　3．剩余平方和　4．回归平方和　5．直线相关

6．零相关　7．相关系数　8．决定系数　9．曲线直线化　10.秩相关

（五）是非题

1． 剩余平方和 $SS_{剩1}=SS_{剩2}$，则 r_1 必然等于 r_2。

2． 直线回归反映两变量间的依存关系，而直线相关反映两变量间的相互直线关系。

3． 两变量关系越密切 r 值越大。

（四）简答题

1． 用什么方法考察回归直线图示是否正确？

2． 剩余标准差的意义和用途？

3． 某资料 $n=100$，X 与 Y 的相关系数为 $r=0.1$，可否认为 X 与 Y 有较密切的相关关系？

4． r 与 r_s 的应用条件有何不同？

5． 应用直线回归和相关分析时应注意哪些问题？

6． 举例说明如何用直线回归方程进行预测和控制？

7． 直线回归分析时怎样确定因变量与自变量？

（五）计算题

1．10 名 20 岁男青年身高与前臂长的数据见表 10-2。

⑴计算相关系数并对 $\rho=0$ 进行假设检验；

⑵计算总体 ρ 的 95% 可信区间。

表 10-2　10 名 20 岁男青年身高与前臂长

身　高（cm）	170	173	160	155	173	188	178	183	180	165
前臂长（cm）	45	42	44	41	47	50	47	46	49	43

2． 某单位研究代乳粉营养价值时，用大白鼠作实验，得到大白鼠进食量和增加体重的数据见表 10-3。

⑴此资料有无可疑的异常点？

⑵求直线回归方程并对回归系数作假设检验。

⑶试估计进食量为 900g 时，大白鼠的体重平均增加多少，计算其 95%的可信区间，并说明其含义。

⑷求进食量为900g 时，个体 Y 值的95%容许区间，并解释其意义。

表 10-3　八只大白鼠的进食量和体重增加量

鼠号	1	2	3	4	5	6	7	8
进食量（g）	800	780	720	867	690	787	934	750
增量（g）	185	158	130	180	134	167	186	133

3． 某省卫生防疫站对八个城市进行肺癌死亡回顾调查，并对大气中苯并（a）芘进行监测，结果如下，试检验两者有无相关？

表 10-4　八个城市的肺癌标化死亡率和大气中苯并（a）芘浓度

城市编号	1	2	3	4	5	6	7	8
肺癌标化死亡率（1/10 万）	5.60	18.50	16.23	11.40	13.80	8.13	18.00	12.10
苯并（a）芘（$\mu g/100m^3$）	0.05	1.17	1.05	0.10	0.75	0.50	0.65	1.20

4． 就下表资料分析血小板和出血症的关系。

表 10-5　12 例病人的血小板浓度和出血症的关系

病例号	1	2	3	4	5	6	7	8	9	10	11	12
血小板数（$10^9/L$）	120	130	160	310	420	540	740	1060	1260	1230	1440	2000
出血症状	++	+++	±	−	+	+	−	−	−	−	++	−

五、习题答题要点

（十五）单项选择题

1.D　2.C　3.C　4.B　5.C　6.C　7.D　8.B　9.D　10.D

（十六）名词解释

1． 直线回归（linear regression）建立一个描述应变量依自变量变化而变化的直线方程，并要求各点与该直线纵向距离的平方和为最小。直线回归是回归分析中最基本、最简单的一种，故又称简单回归（simple regression）。

2． 回归系数（regression coefficient）即直线的斜率(slope)，在直线回归方程中用 b 表示，b 的统计意义为 X 每增（减）一个单位时，Y 平均改变 b 个单位。

3． 剩余平方和（residual sum of squares），$SS_{剩}$ 即 $\sum\left(Y-\hat{Y}\right)^2$，它反映 X 对 Y 的线性影响之外的一切因素对 Y 的变异的作用，也就是在总平方和中无法用 X 解释的部分。在散点图中，

各实测点离回归直线越近，$\sum\left(Y-\hat{Y}\right)^2$ 也就越小，说明直线回归的估计误差越小。

4． 回归平方和（regression sum of squares），$SS_{回}$即$\sum\left(\hat{Y}-\overline{Y}\right)^2$，它反映由于 X 与 Y 的直线关系而使 Y 的总变异所减小的部分，也就是在总平方和中可以用 X 解释的部分。回归平方和越大，说明回归效果越好。

5． 直线相关（linear correlation）又称简单相关（simple correlation），用于双变量正态分布资料。有正相关、负相关和零相关等关系。直线相关的性质可由散点图直观的说明。

6． 零相关（zerro correlation）是指两变量间没有直线相关关系。

29． 相关系数又称积差相关系数（coefficient of product-moment correlation），以符号 r 表示样本相关系数，ρ表示总体相关系数。它是说明具有直线关系的两个变量间，相关关系的密切程度与相关方向的指标。

30． 决定系数（coefficient of determination）即 r 的平方，$r^2=\frac{l_{XY}^2}{l_{XX}l_{YY}}=\frac{l_{XY}^2/l_{XX}}{l_{YY}}=\frac{SS_{回}}{SS_{总}}$，说明当 $SS_{总}$固定不变时，回归平方和的大小决定了 r 平方的大小。回归平方和越接近总平方和，则 r 平方值越接近 1。

31． 曲线直线化（rectification）是曲线拟合的重要手段之一。对于某些非线性的资料可以通过简单的变量变换使之直线化，用直线回归分析方法来分析。

14. 秩相关又称等级相关（rank correlation），是用双变量等级数据作直线相关分析，适用于下列资料：⑴不服从双变量正态分布而不宜作积差相关分析；⑵总体分布型未知；⑶用等级表示的原始数据。

（三）是非题

1．错。两样本剩余平方和 $SS_{剩1}=SS_{剩2}$，但两样本总平方和 $SS_{总}$及回归平方和 $SS_{回}$不一定相等，故两样本相关系数 r_1 与 r_2 不一定相等。

2．正确。

3．错。相关系数 r 有正负之分，其值为 $-1\le r\le 1$，在总体相关系数不为零，即两变量确有直线关系前提下，r 绝对值愈接近 1，两个变量间的直线相关愈密切；愈接近 0，相关愈不密切。

（四）简答题

1．用以下三种方法判定：

⑴直线必须通过点（$\overline{X}$，$\overline{Y}$）。

⑵若纵坐标、横坐标无折断号时，将此线左端延长与纵轴相交，交点的纵坐标必等于截距 a。

⑶直线是否在自变量 X 的实测范围内。

2．剩余标准差用 $s_{Y.X}$ 表示：$s_{Y.X}=\sqrt{SS_{剩}/(n-2)}=\sqrt{\sum\left(Y-\hat{Y}\right)^2/(n-2)}$

其意义是指当 X 对 Y 的影响被扣除后，Y 方面仍有变异。这部分变异与 X 无关，纯属抽样变异。故 $s_{Y.X}$是用来反映 Y的剩余变异的，即不考虑 X 以后 Y 本身的随机变异。剩余标准差可用于：

⑴估计回归系数 b 的标准误，$s_b=s_{Y.X}/\sqrt{l_{XX}}$，进行回归系数的区间估计和假设检验。

⑵估计总体中当 X 为某一定值时，估计值 $\hat{Y}$ 的标准误。$s_{\hat{Y}}=s_{Y.X}\sqrt{1/n+(X-\overline{X})^2/\sum(X-\overline{X})^2}$

并可计算 $\hat{Y}$ 的可信区间，$s_{Y.X}$ 可作为预报精度的指标。

⑶估计总体中当 X 为某一定值时，个体 Y 值的标准差。

批注：考虑 b=0 时，y 估计值是相等的，但此时仍然有剩余平方和存在；y 的估计值不相等，讲的恰好是回归平方和，因为此时估计值与 y 的均数存在离差。

$s_{\hat{Y}}=s_{Y.X}\sqrt{1/n+(X-\overline{X})^2/\sum(X-\overline{X})^2}$，并计算个体 Y 值的容许区间。

3．n=100，r=0.1 时，对相关系数进行 t 检验，按检验水准 α=0.05，拒绝 $H_0(\rho=0)$，接受 $H_1(\rho\neq0)$，认为两变量有相关关系，但决定系数 r^2=0.1^2=0.01，表示回归平方和在总平方和中仅占 1%，说明两变量间的相关关系实际意义不大。

4．积差相关系数 r 用于描述双变量正态分布资料的相关关系。等级相关系数 r_s 适用于下列资料：

⑴不服从双变量正态分布而不宜作积差相关分析的资料；

⑵总体分布型未知的资料；

⑶原始资料是用等级表示的资料。

5．注意以下五个问题

⑴作回归分析和相关分析时要有实际意义，不能把毫无关联的两种现象作回归、相关分析，必须对两种现象间的内在联系有所认识。

⑵在进行回归分析和相关分析之前，应绘制散点图。但观察点的分布有直线趋势时，才适宜作回归、相关分析。如果散点图呈明显曲线趋势，应使之直线化再行分析。散点图还能提示资料有无可疑异常点。

⑶直线回归方程的应用范围一般以自变量的取值范围为限。若无充分理由证明超过自变量取值范围外还是直线，应避免外延。

⑷双变量的小样本经 t 检验只能推断两变量间有无直线关系，而不能推断相关的紧密程度，要推断相关的紧密程度，样本含量必须很大。

⑸相关或回归关系不一定是因果关系，也可能是伴随关系，有相关或回归关系不能证明事物间确有内在联系。

6．用直线回归方程进行预测和控制的步骤

⑴根据研究目的确定预报因子（X）和预报量（Y），由 X 估计 Y 值，收集资料。

⑵建立预报方程 $\hat{Y}=a+bX$，并进行回归系数假设检验。若 P 小于临界值，则回归方程成立。

⑶根据回归方程在 X 实测范围内对 Y 进行预测，并计算 X 为某定值时，个体 Y 值波动范围（容许区间）。

例如，1~7 岁儿童，X 为年龄，Y 为体重，可根据年龄预测（估计）体重。

统计控制是利用回归方程进行逆估计，如要求因变量 Y 值在一定范围内波动，可以通过控制自变量 X 的取值来实现。步骤同前。

例如，针刺哑门穴，进针深度 Y 与颈围 X 间存在直线关系，可根据 X 取值达到控制 Y 的目的。

7．Ⅰ型回归中，X 为精密测量和严格控制的变量，Y 为正态变量。Ⅱ型回归中，X、Y 均为服从正态分布的随机变量，可计算两个回归方程。何者为 X，何者为 Y，根据研究目的确定。例如，测得某一人群的身高和体重两变量，若目的只是由身高估计体重，则确定 X 为身高，Y 为体重。

（五）计算题

1．由原始数据及散点图的初步分析（图 10-1），估计本资料有直线趋势。

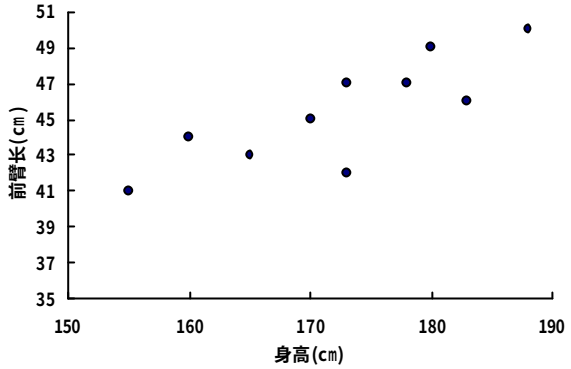


图10-1 10名20岁男青年身高与前臂长散点图

（1）计算相关系数

$\sum X = 1725, \sum X^2 = 298525$，$\overline{X} = 172.5$

$\sum Y = 454, \sum Y^2 = 20690$，$\overline{Y} = 45.4, \sum XY = 78541$

$$l_{XX} = \sum X^2 - \left(\sum X\right)^2 / n = 298525 - 1725^2/10 = 962.5$$

$$l_{YY} = \sum Y^2 - \left(\sum Y\right)^2 / n = 20690 - 454^2/10 = 78.4$$

$$l_{XY} = \sum XY - \left(\sum X\right)\left(\sum Y\right)/n = 78541 - 1725 \times 454/10 = 226$$

$$r = \frac{l_{XY}}{\sqrt{l_{XX} l_{YY}}} = \frac{226}{\sqrt{962.5 \times 78.4}} = 0.8227$$

与 ρ=0 进行假设检验。

H_0：ρ=0，即身高与前臂长间无直线相关关系

H_1：$\rho \neq 0$，即身高与前臂长间有直线相关关系

$$t = \frac{r-0}{s_r} = \frac{r}{\sqrt{\left(1-r^2\right)/(n-2)}} = \frac{0.8227 \times \sqrt{10-2}}{\sqrt{1-0.8227^2}} = 4.09$$

α=0.05

$n = n - 2 = 10 - 2 = 8$，查 t 界值表，得 $0.002 < P < 0.005$，按 α=0.05 水准拒绝 H_0，接受 H_1，故可认为 20 岁男青年身高与前臂长呈正直线相关。

⑵ 算总体 ρ 的 95%可信区间。

对 r 作 z 变换：

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\ln\left(\frac{1+0.8227}{1-0.8227}\right) = 1.1651$$

或，$z = \tanh^{-1} 0.8227 = 1.1651$

z 的 95%可信区间：

$$\left(z-u_{0.05}/\sqrt{n-3}\text{，}\ z+u_{0.05}/\sqrt{n-3}\right)$$
$$=\left(1.1651-1.96/\sqrt{10-3}\text{,}1.1651+1.96/\sqrt{10-3}\right)$$
$$=\left(0.4243\text{,}1.9059\right)$$

按 $r=tanhz$ **对** z **作反变换，得** 20 **岁男青年身高与与前臂长总体相关系数的** 95%**可信区间为（**0.4005，0.9567**）。**

2．由原始数据及散点图初步分析（图 10-2**），估本资料有直线趋势，故作下列计算。**

ΣX=6328，ΣX^2=5048814，$\overline{X}=791$

ΣY=1273，ΣY^2=206619，$\overline{Y}=159.125$ ，ΣXY=1018263

$$l_{XX}=\sum X^2-(\sum X)^2/n=5048814-6328^2/8=43366$$
$$l_{YY}=\sum Y^2-(\sum Y)^2/n=206619-1273^2/8=4052.875$$
$$l_{XY}=\sum XY-(\sum X)(\sum Y)/n=1018263-6328\times1273/8=11320$$
$$b=\frac{l_{XY}}{l_{XX}}=\frac{11320}{43366}=0.261$$
$$a=\overline{Y}-b\overline{X}=159.125-0.261\times791=-47.326$$

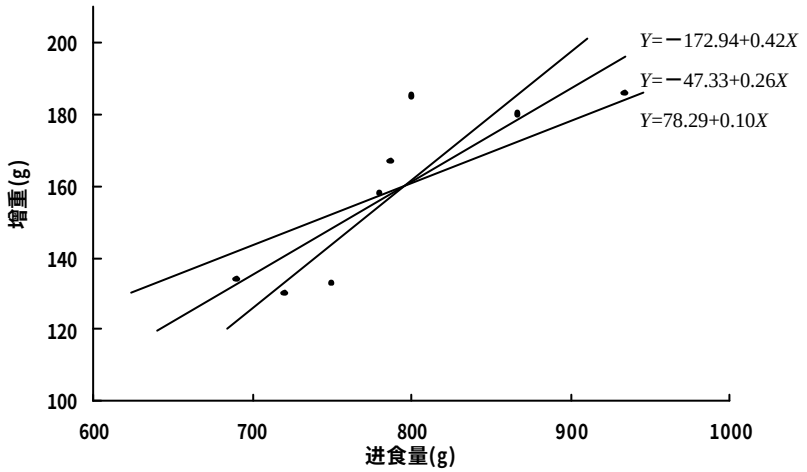


图 10-2 **大白鼠的进食量与增加体重散点图**

（1**）回归系数假设检验：**

H_0：$\beta=0$，**即进食量与增重之间无直线关系**

H_1：$\beta\neq0$，**即进食量与增重之间有直线关系**

$\alpha=0.05$

$SS_{总}=l_{YY}=4052.875$

$SS_{回}=l_{XY}^2/l_{XX}=11320^2/43366=2954.905$

$SS_{剩}=SS_{总}-SS_{回}=4052.875-2954.905=1097.97$

① **方差分析，见表** 10-6**。**

表 10-6 方差分析表

变异来源	*SS*	υ	*MS*	*F*
总变异	4052.875	7		
回归	2954.905	1	2954.905	16.147
剩余	1097.970	6	182.995	

计算得 F=16.147，查 F 界值表，得 P<0.01，按 α=0.05 水准，拒绝 H_0，接受 H_1，可认为大白鼠的进食量与增加体重间有直线关系。

② t 检验：

H_0：β=0，即进食量与增重之间无直线关系

H_1：$\beta\neq0$，即进食量与增重之间有直线关系

α=0.05

$$SS_{总}=l_{YY}=4052.875$$

$$SS_{回}=l_{XY}^2/l_{XX}=11320^2/43366=2954.905$$

$$SS_{剩}=SS_{总}-SS_{回}=4052.875-2954.905=1097.97$$

$$s_{Y.X}=\sqrt{SS_{剩}/(n-2)}=\sqrt{1097.97/(8-2)}=13.5276$$

$$t=\frac{b-0}{s_b}=\frac{b}{s_{Y.X}/\sqrt{l_{XX}}}=\frac{0.261}{13.5276/\sqrt{43366}}=4.018$$

按 υ=6，查 t 界值表，得 0.01>P>0.05，按 α=0.05 水准，拒绝 H_0，接受 H_1，结论同上。本题 $\sqrt{F}=\sqrt{16.147}=4.018=t$

故可用直线回归方程 $\hat{Y}=a+bX=-47.326+0.261X$ 来描述大白鼠的进食量与增加体重的关系。

异常点即对应于残差（$Y-\hat{Y}$）绝对值特大的观测数据见表 10-7。

表 10-7 残差的计算

序号	*X*	*Y*	$\hat{Y}$	$Y-\hat{Y}$
1	800	185	161.474	23.526
2	780	158	156.254	1.746
3	720	130	140.594	−10.594
4	867	180	178.961	1.039
5	690	134	132.764	1.236
6	787	167	158.081	8.919
7	934	186	196.448	−10.448
8	750	133	148.424	−15.424

由散点图及残差分析，第一号点（X=800，Y=185）为可疑的异常点。

⑵根据以上的计算结果，进一步求其总体回归系数的 95%可信区间。绘制回归直线并图示回归系数的 95%可信区间。

总体回归系数 β 的 95%可信区间：

（$b-t_{0.05(n-2)}S_b$，$b+t_{0.05(n-2)}S_b$）

=（0.261－2.447×13.5107／$\sqrt{43366}$ ，0.261＋2.447×13.5107／$\sqrt{43366}$ ）

=（0.1022，0.4198）

取 X_1=690，代入回归方程 $\hat{Y}$ =－47.326+0.261X，得 Y_1=132.76；X_2=934，Y_2=196.45。在图上确定（690，132.76）和（934，196.45）两个点，以直线连接即得回归直线的图形见图 10-2。

按回归系数的 95%可信区间下限和上限分别代入 $a=\overline{Y}-b\overline{X}$，得 a_1=78.285，a_2=－172.937。回归系数的 95%可信区间上、下限对应的两条直线，即图 10-2 中两条回归直线，回归方程为：

$\hat{Y}$ =78.285+0.1022X，$\hat{Y}$ =－172.937+0.4198X

⑶估计进食量为 900g 时，大白鼠的体重平均增加多少，计算其 95%的可信区间，并说明其含义。

$$s_{\hat{Y}}=s_{Y.X}\sqrt{1/n+(X-\overline{X})^2/\sqrt{(X-\overline{X})^2}}$$

$$=13.5276\sqrt{1/8+(900-791)^2/43366}=8.5446$$

当 X=900 时，$m_{\hat{Y}}$ 的 95%可信区间：

（$\hat{Y}-t_{0.05(6)}\,s_{\hat{Y}}$，$\hat{Y}+t_{0.05(6)}\,s_{\hat{Y}}$）

=（187.574－2.447×8.5446，187.574＋2.447×8.5446）=（166.67，208.48）

即总体中，进食量为 900g 时，大白鼠的体重平均增加 187.574g，其 95%的可信区间为 166.67~208.48g。

其含义为：当进食量为 900g 时，相应的平均增重服从一个正态分布（此正态分布的样本均数估计值为 187.574g），如果从此正态分布中重复抽样 100 次，这 100 个可信区间中理论上将有 95 个区间包含真正的总体均数（虽然这个总体均数真值是未知的）。

⑷求进食量为 900g 时，个体 Y 值的 95%容许区间，并解释其意义。

$$s_{Y}=s_{Y.X}\sqrt{1+1/n+(X-\overline{X})^2/\sqrt{(X-\overline{X})^2}}$$

$$=13.5276\sqrt{1+1/8+(900-791)^2/43366}=16.0002$$

当 X=900 时，$\hat{Y}$ =－47.326+0.261X=187.574，个体 Y 值的 95%容许区间：

（$\hat{Y}-t_{0.05(6)}S_Y$，$\hat{Y}+t_{0.05(6)}S_Y$）

=（187.574－2.447×16.0002，187.574＋2.447×16.0002）=（148.42，226.73）

即估计总体中，进食量为 900g 时，有 95%的大白鼠增加体重在 148.42~226.73g 范围内。

3．本题资料不服从双变量正态分布，宜计算等级相关系数。计算过程见表 10-8

表 10-8 八个城市的肺癌标化死亡率和大气中苯并（a）芘的相关分析

城市编号	肺癌标化死亡率（1/10 万） X	等级	苯并（a）芘 Y	等级	d	d^2
(1)	(2)	(3)	(4)	(5)	(6)=(3)－(5)	(7)
1	5.60	1	0.05	1	0	0
2	18.50	8	1.17	7	1	1
3	16.23	6	1.05	6	0	0
4	11.40	3	0.10	2	1	1

5	13.80	5	0.75	5	0	0
6	8.13	2	0.50	3	−1	1
7	18.00	7	0.65	4	3	9
8	12.10	4	1.20	8	4	16
						Σd^2=28

H_0：ρ_s＝0，即肺癌标化死亡率和大气中苯并（a）芘无相关关系

H_1：ρ_s≠0，即肺癌标化死亡率和大气中苯并（a）芘有相关关系

α＝0.05

由上计算表，$r_s=1-6\Sigma d^2/[n(n^2-1)]=1-6\times 28/[8\times(8^2-1)]=0.6667$

查 r_s 界值表，得 $0.10>P>0.05$，按 α＝0.05 水准，不拒绝 H_0，尚不能认为肺癌标化死亡率和大气中的苯并（a）芘有相关关系。

4．本题资料不服从双变量正态分布，宜计算等级相关系数。计算过程见表 10-9。

表 10-9 血小板数与出血症状的等级相关分析

	血小板数（×10^9/L）		出血症状			
病例号	X	等级	Y	等级	d	d^2
(1)	(2)	(3)	(4)	(5)	(6)=(3)−(5)	(7)
1	120	1	++	10.5	−9.5	90.25
2	130	2	+++	12.5	−10.0	100.00
3	160	3	±	7.0	−4.0	16.00
4	310	4	−	3.5	0.5	0.25
5	420	5	+	8.5	−3.5	12.25
6	540	6	+	8.5	−2.5	6.25
7	740	7	−	3.5	3.5	12.25
8	1060	8	−	3.5	4.5	20.25
9	1260	10	−	3.5	6.5	42.25
10	1230	9	−	3.5	5.5	30.25
11	1440	11	++	10.5	0.5	0.25
12	2000	12	−	3.5	8.5	72.25
						Σd^2=402.5

H_0：ρ_s＝0，即血小板数与出血症状无相关关系

H_1：ρ_s≠0，即血小板数与出血症状有相关关系

α＝0.05

因出血症状 Y 中，相同秩次较多，需计算校正 r_s 值 r_s'。

T_X=0

$T_Y=\Sigma(t^3-t)/12=[(6^3-6)+(2^3-2)+(2^3-2)]/12=18.5$

$$r_s' = \frac{\left[\left(n^3-n\right)/6\right]-\left(T_X+T_Y\right)-\sum d^2}{\sqrt{\left[\left(n^3-n\right)/6\right]-2T_X}\sqrt{\left[\left(n^3-n\right)/6\right]-2T_Y}}$$

$$= \frac{\left[\left(12^3-12\right)/6\right]-\left(0+18.5\right)-402.5}{\sqrt{\left[\left(12^3-12\right)/6\right]-0}\sqrt{\left[\left(12^3-12\right)/6\right]-2\times18.5}}$$

$$= -0.5095$$

查 r_s 界值表，得 $0.10>P>0.05$**，按 $\alpha=0.05$ 水准，不拒绝 H_0，尚不能认为血小板数与出血症状有相关关系。**

（王彤　万毅）

第十一章　多元线性回归与 logistic 回归

一、教学大纲要求

（一）掌握内容

1．多元线性回归分析的概念：多元线性回归、偏回归系数、残差。

2．多元线性回归的分析步骤：多元线性回归中偏回归系数及常数项的求法、多元线性回归的应用。

3．多元线性回归分析中的假设检验：建立假设、计算检验统计量、确定 P 值下结论。

4．logistic 回归模型结构：模型结构、发病概率比数、比数比。

5．logistic 回归参数估计方法。

6．logistic 回归筛选自变量：似然比检验统计量的计算公式；筛选自变量的方法。

（二）熟悉内容

常用统计软件（SPSS 及 SAS）多元线性回归分析方法：数据准备、操作步骤与结果输出。

（三）了解内容

标准化偏回归系数的解释意义。

二、教学内容精要

(一) 多元线性回归分析的概念

将直线回归分析方法加以推广，用回归方程定量地刻画一个应变量 Y 与多个自变量 X 间的线形依存关系，称为多元线形回归（multiple linear regression），简称多元回归（multiple regression）

基本形式：

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k$$

式中 $\hat{Y}$ 为各自变量取某定值条件下应变量均数的估计值，X_1，X_2，…，X_k 为自变量，k 为自变量个数，b_0 为回归方程常数项，也称为截距，其意义同直线回归，b_1，b_2，…，b_k 称为偏回归系数（partial regression coefficient**），b_j 表示在除 X_j 以外的自变量固定条件下，X_j 每改变一个单位后 Y 的平均改变量。**

(二) 多元线性回归的分析步骤

$\hat{Y}$ 是与一组自变量 X_1，X_2，…，X_k 相对应的变量 Y 的平均估计值。

多元回归方程中的回归系数 b_1，b_2，…，b_k 可用最小二乘法求得，也就是求出能使估计值 $\hat{Y}$ 和实际观察值 Y 的残差平方和 $\sum e_i^2 = \sum (Y - \hat{Y})^2$ 为最小值的一组回归系数 b_1，b_2，…，b_k 值。根据以上要求，用数学方法可以得出求回归系数 b_1，b_2，…，b_k 的下列正规方程组（normal equation**）：**

$$\begin{cases} b_1 l_{11} + b_2 l_{12} + \Lambda + b_k l_{1k} = l_{1y} \\ b_1 l_{21} + b_2 l_{22} + \Lambda + b_k l_{2k} = l_{2y} \\ \\ b_1 l_{k1} + b_2 l_{k2} + \Lambda + b_k l_{kk} = l_{ky} \end{cases}$$

式中

$$l_{ij} = l_{ji} = \sum (X_i - \overline{X}_i)(X_j - \overline{X}_j) = \sum X_i X_j - \frac{(\sum X_i)(\sum X_j)}{n}$$

$$l_{iy} = \sum (X_i - \overline{X}_i)(Y - \overline{Y}) = \sum X_i Y - \frac{(\sum X_i)(\sum Y)}{n}$$

常数项b_0可用下式求出：

$$b_0 = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2 - \Lambda - b_k \overline{X}_k$$

（三）多元线性回归分析中的假设检验

在算得各回归系数并建立回归方程后，还应对此多元回归方程作假设检验，判断自变量X_1，X_2，…，X_k是否与Y真有线性依存关系，也就是检验无效假设H_0（$b_1 = b_2 = b_3 = \mathrm{L} = b_k = 0$），备选假设$H_1$为各$b_j$值不全等于0或全不等于0。

检验时常用统计量F

$$F = \frac{MS_{回归}}{MS_{误差}} = \frac{l_{回归}/k}{l_{误差}/(n-k-1)}$$

式中n为个体数，k为自变量的个数。

式中

$$l_{回归} = b_1 l_{1y} + b_2 l_{2y} + \Lambda + b_k l_{ky}$$

$$l_{误差} = l_{总} - l_{回归}$$

$$l_{总} = \sum \left(Y - \overline{Y}\right)^2 = l_{yy}$$

(四) logistic回归模型结构

设X_1, X_2, Λ, X_k为一组自变量，Y为应变量。当Y是阳性反应时，记为Y=1；当Y是阴性反应时，记为Y=0。用P表示发生阳性反应的概率；用Q表示发生阴性反应的概率，显然$P+Q$=1。

Logistic回归模型为：

$$P = \frac{e^{b_0 + b_1 X_1 + b_2 X_2 + \Lambda + b_k X_k}}{1 + e^{b_0 + b_1 X_1 + b_2 X_2 + \Lambda + b_k X_k}}$$

同时可以写成：

$$Q = \frac{1}{1 + e^{b_0 + b_1 X_1 + b_2 X_2 + \Lambda + b_k X_k}}$$

式中b_0是常数项；$b_j (j = 1, 2, \mathrm{L}, k)$是与研究因素$X_j$有关的参数，称为偏回归系数。

事件发生的概率P与$b\,x$之间呈曲线关系，当$b\,x$在$(-\infty, \infty)$之间变化时，P或Q在（0，1）之间变化。

若有n例观察对象，第i名观察对象在自变量$X_{i1}, X_{i2}, \Lambda, X_{ik}$作用下的应变量为$Y_i$，阳性

反应记为Y_i=1，否则Y_i=0。相应地用P_i表示其发生阳性反应的概率；用Q_i表示其发生阴性反应的概率，仍然有P_i+Q_i=1。P_i和Q_i的计算如下：

$$P_i = \frac{e^{b_0+b_1X_{i1}+b_2X_{i2}+\mathrm{L}\ +b_kX_{ik}}}{1+e^{b_0+b_1X_{i1}+b_2X_{i2}+\mathrm{L}\ +b_kX_{ik}}}$$

$$Q_i = \frac{1}{1+e^{b_0+b_1X_{i1}+b_2X_{i2}+\mathrm{L}\ +b_kX_{ik}}}$$

这样，第i个观察对象的发病概率比数（odds）为P_i/Q_i，第l个观察对象的发病概率比数为P_l/Q_l，而这两个观察对象的发病概率比数之比值便称为比数比OR（odds ratio)。对比数比取自然对数得到关系式：

$$\ln\left(\frac{P_i/Q_i}{P_l/Q_l}\right)= b_1(X_{i1}-X_{l1})+b_2(X_{i2}-X_{l2})+\Lambda\ +b_k(X_{ik}-X_{lk})$$

等式左边是比数比的自然对数，等式右边的$\left(X_{ij}-X_{lj}\right)\left(j=1,2,\Lambda\ ,\ k\right)$是同一因素$X_i$的不同暴露水平$X_{ij}$与$X_{lj}$之差。$b_j$的流行病学意义是在其它自变量固定不变的情况下，自变量X_j的暴露水平每改变一个测量单位时所引起的比数比的自然对数改变量。或者说，在其他自变量固定不变的情况下，当自变量X_j的水平每增加一个测量单位时所引起的比数比为增加前的e^{b_j}倍。同多元线性回归一样，在比较暴露因素对反应变量相对贡献的大小时，由于各自变量的取值单位不同，也不能用偏回归系数的大小作比较，而须用标准化偏回归系数来做比较。标准化偏回归系数值的大小，直接反映了其相应的暴露因素对应变量的相对贡献的大小。标准化偏回归系数的计算，可利用有关统计软件在计算机上解决。

（五）logistic回归参数估计

由于logistic回归是一种概率模型,通常用最大似然估计法(maximum likelihood estimate)求解模型中参数b_j的估计值$b_j\ (j=1,2,\mathrm{L}\ ,\ k)$。

Y为在$X_1,X_2,\Lambda\ ,X_k$作用下的阳性事件（或疾病）发生的指示变量。其赋值为：

$$Y_i=\begin{cases}1，第i个观察对象出现阳性反应\\0，第i个观察对象出现阴性反应\end{cases}$$

第i个观察对象对似然函数的贡献量为：

$$l_i = P_i^{Y_i}Q_i^{1-Y_i}$$

当各事件是独立发生时，则n个观察对象所构成的似然函数L是每个观察对象的似然函数贡献量的乘积，即

$$L=\prod_{i=1}^{n}l_i=\prod_{i=1}^{n}P_i^{Y_i}Q_i^{1-Y_i}$$

式中$\prod$为i从1到n的连乘积。

依最大似然估计法的原理，使得L达到最大时的参数值即为所求的参数估计值，计算时通常是将该似然函数取自然对数（称为对数似然函数）后，用Newton—Raphson迭代算法求

解参数估计值b_j（$j=1,2,\Lambda$，k）。

（六）logistic 回归筛选自变量

在 logistic 回归中，筛选自变量的方法有似然比检验（likelihood ratiotest）、计分检验(score test)、Wald 检验(Wald test)三种。其中似然比检验较为常用，

用Λ表示似然比检验统计量，计算公式为：

$$\Lambda = 2\ln\left(L^{'}/L\right) = 2(\ln L^{'} - \ln L)$$

式中 ln 为自然对数的符号，L 为方程中包含$m(m<k)$个自变量的似然函数值，$L^{'}$为在方程中包含原m个自变量的基础上再加入 1 个新自变量X_j后的似然函数值。在无效假设H_0条件下，统计量Λ服从自由度为 1 的c^2分布。当$\Lambda \geq c^2_{a(1)}$时，则在a水平上拒绝无效假设，即认为X_j对回归方程的贡献具有统计学意义，应将X_j引入到回归方程中；否则，不应加入。逆向进行即可剔除自变量。

三、典型试题分析

（一）单项选择题

1．多元线性回归分析中，反映回归平方和在应变量Y的总离均差平方和中所占比重的统计量是（　　）。

D．复相关系数

E．偏相关系数

F．偏回归系数

D．确定系数

答案：D

[评析]　本题考点：多元线性回归中的几个概念的理解。

多元线性回归中的偏回归系数（multiple linear regression）表示在其它自变量固定不变的情况下，自变量X_j每改变一个单位时，单独引起应变量Y的平均改变量。确定系数（coefficient of determination）表示回归平方和$SS_{回归}$占总离均差平方和$SS_{总}$的比例，简记为R^2。即 $R^2 = SS_{回归}/SS_{总}$。确定系数的平方根即R称为复相关系数（multiple correlation coefficient），它表示p个自变量共同对应变量线性相关的密切程度，它不取负值，即$0 \leq R \leq 1$。

2．Logistic回归分析适用于应变量为（　　）。

A．分类值的资料　　B．连续型的计量资料

C．正态分布资料　　D．一般资料

答案：A

[评析] 本题考点：logistic回归的概念。

logistic 回归属于概率型回归，可用来分析某类事件发生的概率与自变量之间的关系。适用于应变量为分类值的资料，特别适用于应变量为二项分类的情形。模型中的自变量可以是定性离散值，也可以是计量观测值。

（二）计算题

根据表 11-2 数据，分别用 SPSS 统计软件、SAS 统计软件写出多元线性回归的统计分

析步骤及其简要结果。

表 11-1　某学校 20 名一年级女大学生肺活量及有关变量测量结果

编号	体重 X_1/kg	胸围 X_2 /cm	肩宽 X_3 /cm	肺活量 Y /L
1	50.8	73.2	36.3	2.96
2	49.0	84.1	34.5	3.13
3	42.8	78.3	31.0	1.91
4	55.0	77.1	31.0	2.63
5	45.3	81.7	30.0	2.86
6	45.3	74.8	32.0	1.91
7	51.4	73.7	36.5	2.98
8	53.8	79.4	37.0	3.28
9	49.0	72.6	30.1	252
10	53.9	79.5	37.1	3.27
11	48.8	83.8	33.9	3.10
12	52.6	88.4	38.0	3.28
13	42.7	78.2	30.9	1.92
14	52.5	88.3	38.1	3.27
15	55.1	77.2	31.1	2.64
16	45.2	81.6	30.2	2.85
17	51.4	78.3	36.5	3.16
18	48.7	72.5	30.0	2.51
19	51.3	78.2	36.4	3.15
20	45.8	75.0	32.5	1.94

答案：

SPSS：**数据文件：**“ EXAP11—2．sav ”。 **数据格式：** 4 **列** 20 **行。过程：**

Statistic
Regression
Linear...
Dependent：***Y***
Independent(s)： X_1， X_2， X_3
Method： Enter

结果：

Variables Entered/Removed

Model	Variables Entered	Variables Removed	Method
1	X_3 **（肩宽）**, X_2 **（胸围）**,	.	Enter

	X_1（体重）		

a All requested variables entered.

b Dependent Variable: *Y*（肺活量）

Model Summary

Mode	R	R Square	Adjusted RSquare	Std. Error of the Estimate
1	.846	.715	.662	.2893

a Predictors: (Constant), X_3, X_2, X_1

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3.367	3	1.122	13.413	.000
	Residual	1.339	16	8.368E-02		
	Total	4.706	19			

a Predictors: (Constant), X_3, X_2, X_1

b Dependent Variable: *Y*

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	*t*	Sig.
		B	Std. Error	Beta		
1	(Constant)	-4.676	1.321		-3.541	.003
	X_3	6.036E-02	.021	.474	2.899	.010
	X_2	3.508E-02	.015	.333	2.272	.037
	X_1	5.010E-02	.029	.307	1.735	.102

a Dependent Variable: *Y*

SAS：

数据步

DATA EXAP11—2；INPUT x1 x2 x3 y@ @；

CARDS；

50.8 73.2 36.3 2.96…45.8 75.0 32.5 1.94；

过程步

PROC REG；

MODEL y=x1 x2 x3；

RUN；

结果：

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3.36732	1.12244	13.41	0.0001
Error	16	1.33893	0.08368		
Corrected Total	19	4.70626			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-4.67553	1.32051	-3.54	0.0027
X1	1	0.06036	0.02082	2.90	0.0105
X2	1	0.03508	0.01544	2.27	0.0372
X3	1	0.05010	0.02888	1.73	0.1020

[评析]　本题考点：统计软件关于多元线性回归的分析方法及主要输出结果。

根据 SPSS **或** SAS **的输出结果，可进行以下分析：**

1．**检验** H_0：$b_1 = b_2 = b_3 = 0$ **的方差分析表。** F=13.413，P=0.0001，**拒绝** H_0，**肺活量至少与一个自变量存在线性关系。**

2．**估计偏回归系数** b_1，b_2，b_3，**给出多元线性回归方程**

$\hat{Y} = 4.68 + 0.06X_1 + 0.04X_2 + 0.05X_3$，$R^2$=0.715，$R_a^2$=0.662。

3．**偏回归系数检验，见表** 11-2。

表 11-2　**偏回归系数估计值及其检验**

偏回归系数	估计值	*SE*	*t*	*P*
b_0	-4.675	1.321	-3.54	0.00
b_1	0.060	0.021	2.90	0.01
b_2	0.035	0.015	2.27	0.04
b_3	0.050	0.029	1.73	0.10

四、习　题

（十五）　单项选择题

32．　**可用来进行多元线性回归方程的配合适度检验是：**

A．　c^2 **检验**　　B．　F **检验**

C．　U **检验**　　D．　Ridit **检验**

33．　**在多元回归中，若对某个自变量的值都增加一个常数，则相应的偏回归系数：**

A．　**不变**

B．　**增加相同的常数**

C．　**减少相同的常数**

D．　**增加但数值不定**

34． 在多元回归中，若对某个自变量的值都乘以一个相同的常数 k，则：

B． 该偏回归系数不变

C． 该偏回归系数变为原来的 1/k 倍

D． 所有偏回归系数均发生改变

E． 该偏回归系数改变，但数值不定

35． 作多元回归分析时，若降低进入的 F 界值，则进入方程的变量一般会：

A． 增多　　B． 减少

C． 不变　　D． 可增多也可减少

（二） 名词解释

1.多元线性回归　　2.偏回归系数　　3.复相关系数　　4.确定系数

5.比数　　6.比数比

（三） 简答题

logistic 回归模型中，偏回归系数 b_i 的解释意义是什么？

（四） 计算题

某学者研究在某种营养缺乏状态下儿童的体重（Y，kg）与身高（X_1，cm）、年龄（X_2，岁）的关系获得了 12 名观察对象的观测资料，计算得到如下基本数据：

$\sum X_1 = 1611$，$\sum X_1^2 = 219631$，$\sum X_2 = 106$，$\sum X_2^2 = 976$，$\sum Y = 341$，$\sum Y^2 = 9883$，$\sum X_1 X_2 = 14454$，$\sum X_1 Y = 46439$，$\sum X_2 Y = 3079$。

（1） 请写出求解 $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$ 二元线性回归方程的正规方程组。

（2） 设方程组的解为 $b_0 = 2.114$，$b_1 = 0.135$，$b_2 = 0.923$，请写出回归方程。

（3） 完成下列方差分析表。

表 11-3　12 名儿童体重与身高、年龄回归分析方差分析表

变异来源	v	SS	MS	F
回归				
残差				
总和				

五、习题答案要点

（一） 单项选择题

1． B　　2． A　　3． B　　4． A

（二） 名词解释

1． 用回归方程定量地刻画一个应变量 Y 与多个自变量 X 间的线性依存关系，称为多元线性回归（multiple linear regression），简称多元回归（multiple regression）。

2． 多元线性回归的基本形式为：$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k$　b_1，b_2，…，b_k 称为偏回归系数（partial regression coefficient），b_j 表示在除 X_j 以外的自变量固定条件下，X_j 每改变一个单位后 Y 的平均改变量。

3． 复相关系数 R（coefficient of multiple correlation），R 的大小表示所有自变量与应变量之间线性关系的密切程度。

4． 确定系数（coefficient of determination）简记为R^2，表示回归平方和$SS_{回归}$占总离均差平方和$SS_{总}$的比例，即 $R^2 = SS_{回归} / SS_{总}$。用R^2可定量评价在 y 的总变异中，由 x 变量组建立的线性回归方程所能解释的比例。

5． logistic回归模型为：

$$P = \frac{e^{b_0+b_1X_1+b_2X_2+\Lambda+b_kX_k}}{1+e^{b_0+b_1X_1+b_2X_2+\Lambda+b_kX_k}}$$

同时可以写成：

$$Q = \frac{1}{1+e^{b_0+b_1X_1+b_2X_2+\Lambda+b_kX_k}}$$

第 i 个观察对象的发病概率比数（odds）为 P_i/Q_i，即同一暴露水平下，阳性概率与阴性概率之比值称为比数（odds）。

6． logistic 回归模型中，两个观察对象的发病概率比数之比值称为比数比 OR（odds ratio）。其大小反映了不同暴露水平下，个体发病的相对危险程度。

（三）简答题

答：b_j 的流行病学意义是在其它自变量固定不变的情况下，自变量 X_j 的暴露水平每改变一个测量单位时所引起的比数比的自然对数改变量。或者说，在其他自变量固定不变的情况下，当自变量 X_j 的水平每增加一个测量单位时所引起的比数比为增加前的 e^{b_j} 倍。

（四） 计算题

1．求解 $\hat{Y} = b_0 + b_1X_1 + b_2X_2$ 二元线性回归方程的正规方程组为：

$$\begin{cases} b_1l_{11} + b_2l_{12} = l_{1y} \\ b_1l_{21} + b_2l_{22} = l_{2y} \end{cases}$$

2．当方程组的解为 $b_0 = 2.114$，$b_1 = 0.135$，$b_2 = 0.923$，回归方程为：

$$\hat{Y} = 2.114 + 0.135\,X_1 + 0.923\,X_2$$

3．列方差分析表。

表 11-4　12名儿童体重与身高、年龄回归分析方差分析表

变异来源	v	SS	MS	F
回归	2	151.35	75.675	16.380
残差	9	41.57	4.62	
总和	11	192.92		

（尹平　白玉祥）

第十二章　统计表与统计图

一、教学大纲要求

（一）掌握内容

1．统计表

（1） 统计表的结构。

（2） 统计表的种类。

（3） 编制统计表的注意事项。

2．统计图

（1） 统计图的结构。

（2） 统计图的种类。

（3） 统计图的编制要求。

（二）熟悉内容

常用统计图的绘制方法和注意事项。

（三）了解内容

半对数线图、箱式图、误差线图等的绘制方法和注意事项。

二、教学内容精要

（一）统计表与统计图的概念

将统计资料及其指标以表格形式列出，称为统计表（statistical table)。狭义的统计表只表示统计指标。

统计图(statistical graph)是将统计指标以点的位置、线段的升降、直条的长短或面积的大小等几何图形直观的表示事物间的数量关系。

（二）统计表中应注意的几个问题

1．列表的原则

（1）重点突出，简单明了。

（2）主次分明，层次清楚，符合逻辑。

2．统计表的结构与编制要求

统计表由标题、标目、线条和数字所构成。如下表所示：

顶线

表 号　　标　题

横标目名称	纵标目名称	合计
横标目	数　字	
合 计		

底线

（1） 标题

位于表的上方，概括表的主要内容，一般需注明时间与地点。

（2）标目

有横、纵标目之分，分别说明横行和纵行数字的含义，应做到文字简明，层次清楚。

（3）线条

多采用三条半线，即顶线、底线、纵标目下的横隔线及合计上的半线。忌斜线和竖线。

（4）数字

表内数据一律采用阿拉伯数字。同一指标小数点位数要一致，位次要对齐。表内不应有空项，无数字用“ — ”表示，数字若为零则填“ 0 ”，暂缺项或未记录用“ … ”表示。

（5）备注

不为表的必备内容，如有必要，可在表内用“ * ”号标记，然后在表的下方加以说明。

3．统计表的种类

统计表可分为简单表(simple table)和复合表(combinative table)两种类型。

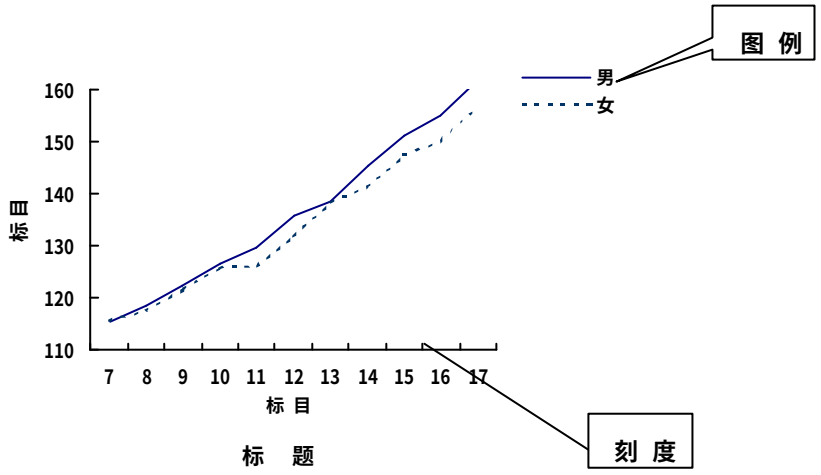
简单表：只按单一特征或标志分组。

复合表：按两个或两个以上主要标志分组,如年龄和性别结合起来分组。

（三）统计图中应注意的几个问题

1．统计图的结构

统计图通常由标题、标目、刻度和图例四部分组成。如下所示：



2．常用统计图的分类

医学统计学中常用的统计图有：条图(bar graph)、线图(line graph)、圆图 (pie graph)、直方图（histogram)、散点图(scatter diagram)和统计地图(statistical map)等。

3．制图的基本要求

(1)根据资料的性质和分析目的，选择合适的图形。

(2)标题扼要说明图的主要内容，位于图的下方，必要时注明时间和地点。

(3)建立在直角坐标系上的统计图，其纵轴尺度自下而上，横轴尺度从左到右，数字一律由小到大，某些图还要求纵轴尺度从 0 开始（如直条图、直方图)。纵横两轴一般应有标目，注明单位。

(4)图的长宽比例（圆图除外）一般以 7:5 或 5:7 为宜。

(5)可用不同的线条或颜色表示不同的事物，但需用图例说明，一般放在图的右上角或图的下方。

4.常用统计图的定义和制图要求，见表 12-1。

表 12-1　常用统计图的定义和制图要求

名　称	定　义	制 图 要 求
条　图	用等宽直条的长短来表示相互独立的各统计指标的数值大小	起点为 0 的等宽直条，条间距相等，按高低顺序排列。
普通线图	适用于连续性资料。用线段的升降来表示一事物随另一事物变化的趋势。	纵横两轴均为算术尺度，相邻两点应以折线相连。图内线条不宜超过 3 条。
半对数线图	用线段的升降来表示一事物随另一事物变化的速度。	横轴为算术尺度，纵轴为对数尺度。余同普通线图。
圆　图	以圆面积表示事物的全部，用扇形面积表示各部分的比重	以圆面积为 100%，将各构成比分别乘以 3.6 度得圆心角度数后再绘扇形面积。通常以 12 点为始边依次绘图。
直方图	用矩形的面积来表示某个连续型变量的频数分布	常以横轴表示连续型变量的组段（要求等距），纵轴表示频数或频率，其尺度从“ 0″开始，各直条间不留空隙。
散点图	以点的密集程度和趋势表示两种事物间的相关关系	绘制方法同线图，只是点与点之间不连接。

三、典型试题分析

1. 指出表 12-2 的缺陷并作改进。

表 12-2　119 例宫颈糜烂冷冻治疗结果　（原表）

	轻度糜烂		中度糜烂		重度糜烂		总计	
	例数	%	例数	%	例数	%	例数	%
治愈	39	32.77	11	9.24	2	1.68	52	43.70
好转	2	1.68	19	15.97	14	11.76	35	29.41
无效	8	6.72	7	5.88	17	14.29	32	26.89
合计	49		37		33		119	

[评析]本题考点：对列表的原则和统计表的结构与编制要求的掌握。

表 12-2 的主要目的在于考察冷冻治疗宫颈糜烂的近期疗效。存在的问题是：标题未突出“近期疗效”这一主要内容；主谓语安排不当且标目重复，如例数和%多处出现；总计意义不明确；线条过多，以致数据隔离，不便比较。改正后见表 12-3。

表 12-3　冷冻治疗宫颈糜烂患者的近期疗效　　（修改表）

糜烂程度	例数	疗效			疗效构成比（%）		
		治愈	好转	无效	治愈	好转	无效
轻　度	49	39	2	8	79.6	4.1	16.3
中　度	37	11	19	7	29.7	51.4	18.9
重　度	33	2	14	17	6.1	42.4	51.5
合　计	119	52	35	32	43.7	29.4	26.9

修改表 12-3 很容易看清楚冷冻治疗宫颈糜烂中治愈、好转、无效在各级糜烂程度中的例数和所占的百分比，同时也可以看出疗效因宫颈糜烂程度不同而异，轻度糜烂者疗效较好，中、重度次之。

2．将下表资料绘成合适的图形。

表 12-4 亚洲国家成人 HIV 感染情况

国家	成人感染率（%）
柬埔寨	2.40
泰国	2.23
缅甸	1.79
印度	0.82
中国	0.06

[评析]本题考点：对各种统计图适用情况的掌握。

分析表 12-4 的资料，得出此资料适合做单式条图，见图 12-1。

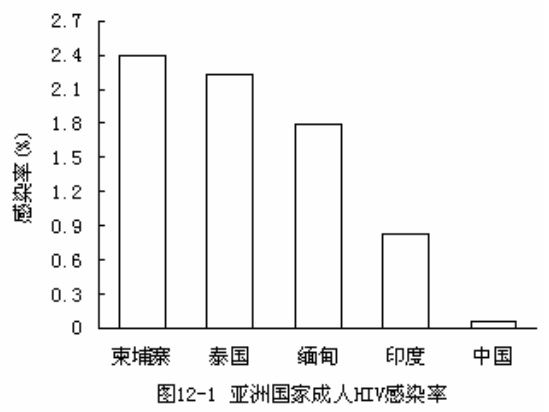


图12-1　亚洲国家成人HIV感染率

3.根据表 12-5 的资料，作图并作简要分析。

表 12-5　某市某年男女学生不同年龄的身高均数(cm)

年龄组（岁）	男	女
17～	115.41	115.51
18～	118.33	117.53
19～	122.16	121.66
10～	129.48	125.94
11～	129.64	131.76
12～	135.50	138.26
13～	138.36	141.17
14～	145.14	147.21
15～	150.84	150.03
16～	154.70	153.06
17～18	161.90	156.63

[评析]本题考点：对统计图的做法与分析知识点的掌握。

绘线图，见图 12-2。

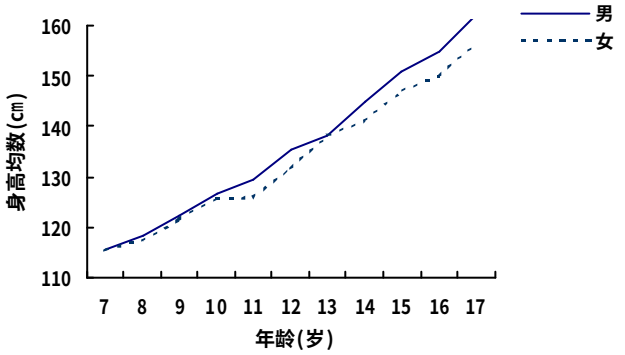


图12-2 某市某年男女学生不同年龄身高均数

由表 12-5 和图 12-2 可见，随着年龄的增加，男女生身高均数均逐渐增高。在 7～10 岁间，男生身高均数略高于女生；而 10～15 岁间，男生身高均数略低于女生；15 岁以上，男生身高均数又超过女生，表现出不同性别儿童生长发育曲线的两交叉现象。

4．根据表 12-6 的资料，做合适的图形并作简述作图步骤。

表 12-6　我国 1998 年性病传播途径分布情况

传播途径	病例数	构成比（%）
非婚姻性接触	413 303	72.1
配偶传播	103 064	18.0
其他传播	57 174	9.9

[评析] 本题考点：圆图的应用。

圆图是用圆的总面积表示事物的全部，用各个扇形的面积表示各个部分的比重，根据资料的性质，此题适用于作圆图。

(1) 先计算各部分的角度　根据公式圆心角（度）=各部分百分比×360°。

(2) 绘制图形　先画出圆形，再借助量角器画出各圆心角。以第一个圆心角从时钟9点或12点处开始，顺时针方向排列。如下图12-3。

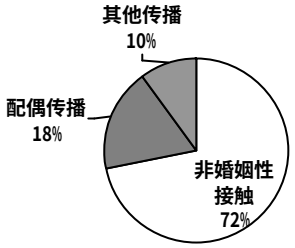


图12-3 我国1998年性病传播途径构成比

5．将表12-7资料中两种疾病发病率的历年变动情况绘制成普通线图及半对数线图，并说明两种图形的不同意义。

表12-7　某地结核病和白喉的死亡率（‰）

年 份	结核病死亡率	白喉死亡率
1949	150.2	20.1
1950	148.0	16.6
1951	141.0	14.0
1952	130.0	11.8
1953	110.4	10.7
1954	98.2	6.5
1955	72.6	3.9
1956	68.0	2.4
1957	54.8	1.3

[评析] 本题考点：半对数线图的应用。

半对数线图是线图的一种特殊形式，在事物数量间相差较大的情况下，通常普通线图难于表达或相互比较两种或两种以上事物的变化速度，此时可采用半对数图来表示。

(1)　普通线图：

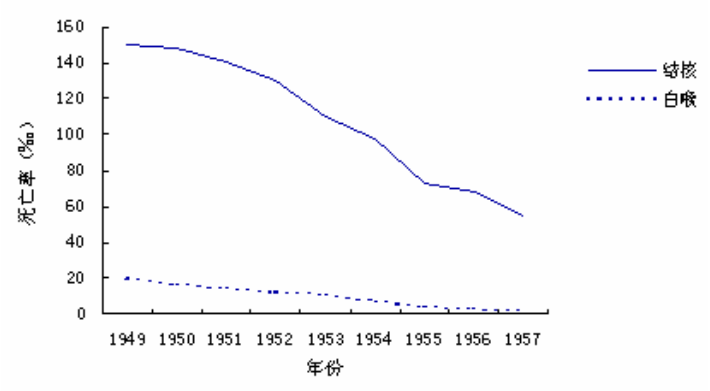


图 12-4 某市 1949-1957 年 15 岁以下儿童结核、白喉死亡率

由纵横两轴均为算术尺度的普通线图 12-4 可见，结核病和白喉死亡率 1949-1957 年均呈下降趋势，给人们的直观感觉是结核病的死亡率下降较快，而白喉死亡率下降较平缓。

（2）半对数线图

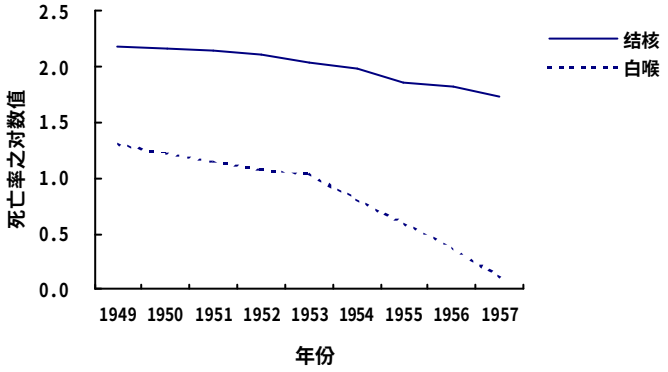


图12-5 某市1949-1957年15岁以下儿童结核、白喉死亡率

由半对数线图可见，结核病的死亡率下降速度始终比较平缓，而白喉死亡率下降速度开始几年和结核病持平，1954 年后下降速度明显加快。

四、习 题

（一）名词解释

1．统计表　　2．统计图

（二）简答题

1．统计表和统计图在表达资料中各有何特殊意义？

2．统计表有哪些要素构成？制表的注意事项有哪些？

3．统计图有哪些要素构成？绘制统计图的注意事项有哪些？

4．为什么半对数线图可以描述发展速度的变化？

（三）列表、制图与分析题

1.某医院对麦芽根糖浆治疗急慢性肝炎 161 例的疗效列表，试作改进。

表 12-8　麦芽根糖浆治疗急慢性肝炎疗效观察

效果 / 总例数	有效						无效	
	小计		近期痊愈		好转			
	例	%	例	%	例	%	例	%
161	108	67.1	70	43.5	38	23.6	53	32.9

2．某地 1952 年和 1972 年三种死因别死亡率下表，试将该资料绘制成统计图并作分析。

表 12-9　某地 1952 年和 1972 年三种死因别死亡率（1/10 万）

死因	1952 年	1972 年
肺结核	165.2	27.4
心脏病	72.5	83.6
恶性肿瘤	57.2	178.2

3．据下例统计资料试作统计图。

表 12-10　某地居民两次粪便蠕虫卵检查结果

	第一次阳性率（%）	第二次阳性率（%）
蛔 虫	91.43	86.39
钩 虫	61.22	31.36
鞭 虫	17.14	16.51

表 12-11　某部队 1997 年各月传染病发病人数

月 份	1	2	3	4	5	6	7	8	9	10	11	12	合计
传染病人数	3	4	7	14	9	14	17	104	58	12	5	2	249

表 12-12　224 例胸膜炎病人的年龄分布

年龄（岁）	各组人数占全部病人的百分比
11～	4.1
16～	13.5
21～	44.6
31～	27.1
41～	8.9
51～	1.8
合　　计	100.0

4．某县防疫站 1972 年开始在城关镇建立“预防接种卡”，使计划免疫得到加强。为说明效果，1975 年 5 月观察了 482 人的锡克试验反应，其中：幼儿园儿童 101 人，阳性 21 人；小学生 145 人，阳性 22 人；中学生 236 人，阳性 15 人。相比起来，1947 年为：幼儿园儿童 144 人，阳性 37 人；小学生 1417 人，阳性 323 人；中学生 359 人，阳性 41 人。试用适当的统计表和统计图描述上述结果，并作简要分析。

（四）是非题

1．一个绘制合理的统计图可直观的反映事物间的正确数量关系。

2．在一个统计表中，如果某处数字为“0”，就填“0”，如果数字暂缺则填“…”，如果该处没有数字，则不填。

3．备注不是统计表的必要组成部分，不必设专栏，必要时，可在表的下方加以说明。

4．散点图是描写原始观察值在各个对比组分布情况的图形，常用于例数不是很多的间断性分组资料的比较。

5．百分条图表示事物各组成部分在总体中所占比重，以长条的全长为 100%，按资料的原始顺序依次进行绘制，其他置于最后。

五、习题答题要点

（一） 名词解释

1．统计表：将统计资料及其指标以表格形式列出，称为统计表（statistical table)。狭义的统计表只表示统计指标。

2．统计图：统计图(statistical graph)是将统计指标用几何图形表达，即以点的位置、线段的升降、直条的长短或面积的大小等形式直观的表示事物间的数量关系。

（二） 简答题

1．统计表可以代替冗长的文字叙述，便于指标的计算、分析和对比，其制作合理与否，对统计分析质量有着重要的影响。

统计图可用点的位置、线段的升降、直条的长短和面积的大小直观地反映分析事物间的数量关系。因统计如对数量表达较粗略，故最好附上相应的统计表。

2．一般说来，统计表由标题、标目、线条、数字四部分构成（有时附有备注)。

编制统计表的注意事项：

(1) 标题概括表的内容，写于表的上方，通常需注明时间与地点。

(2) 标目以横、纵标目分别说明主语与谓语，文字简明，层次清楚。

(3) 线条不宜过多，通常采用三条半线表示，即顶线、底线、纵标目下的横隔线及合计上的半条线 。

(4) 表内一律采用阿拉伯数字。同一指标小数点位数要一致，数次要对齐。表内不留空格。

(5) 备注不要列于表内，如有必要，可在表内用“ * ”号标记，并在表外加以说明。

3．统计图通常由标题、标目、刻度和图例四部分组成。

绘制统计图的注意事项：

(1) 根据资料的性质和分析目的，选择合适的图形。

(2) 标题应扼要的说明图的内容、地点、时间，位于图的下方，一般需注明时间、地点。

(3) 统计图有纵轴和横轴，两轴应有标目，标目应注明单位。纵轴尺度自下而上，横轴尺度从左到右。数字一律由小到大，某些图要求纵轴尺度从 0 开始

(4) 图的长宽比例（除圆图外）一般以 7:5 或 5:7 左右较美观。

(5) 比较不同事物时，可用不同的线条或颜色表示，但需用图例说明，一般放在图的右上角或图下方的适当位置。

半对数线图是以横轴为算术尺度，纵轴为对数尺度绘制而成。它表明数量间比例的动态变化趋势，如速率比 A/B，设 X=A/B,利用对数运算法则，lgX= lgA – lgB,即将纵轴上尺度的倍比关系用对数值之差表示，所以它反映的是 A ，B 两事物现象间相互对比发展速度的变化。

（三） 列表、制图与分析题

1．对表 12-8 进行改进后，见表 12-13。

表 12-13　某医院麦芽根糖浆治疗急慢性肝炎疗效分析

疗　效	例数	疗效构成比（%）
无　效	53	32.92
好　转	38	23.60
近期痊愈	70	43.48
合　计	161	100.00

2． 根据资料性质，将资料绘成复式条图，见图 12-6。

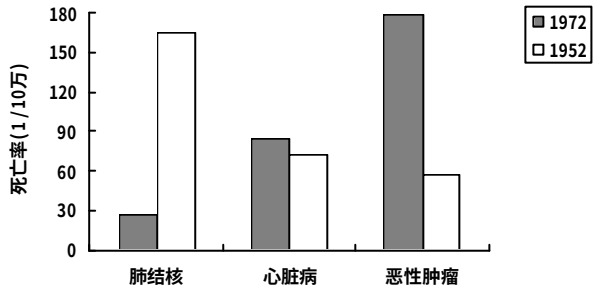


图12-6　某地两年三种死因别死亡率(1/10万)比较

由图可见，1972 年与 1952 年相比肺结核死亡率明显下降；心脏病死亡率两年相比轻微增高；恶性肿瘤死亡率急剧上升，提示不同时期死因别死亡率的变化情况，反映出不同时期疾病防治的重点。

3．表 12-10 绘成直条图，见图 12-7。表 12-11 绘成线图，见图 12-8。表 12-12 将组段改为等距后（见表 12-14），绘成直方图，见图 12-9。

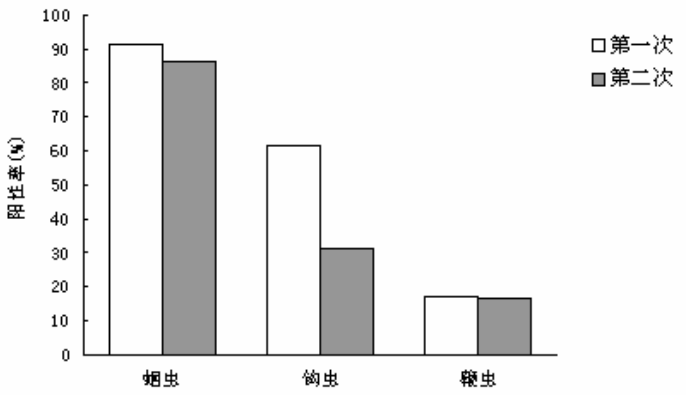


图12-7 某地居民粪便中蠕虫卵两次检查结果

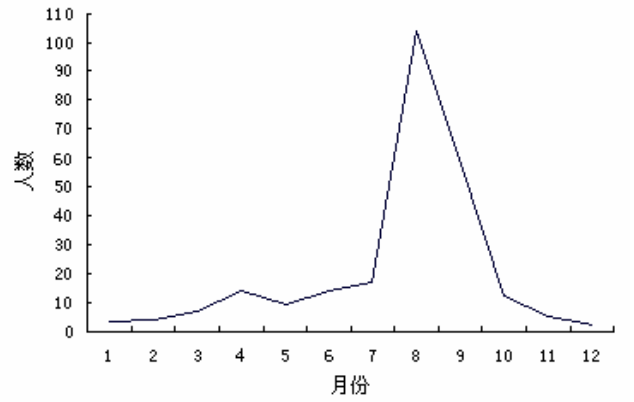


图12-8 某部队1997年每月传染病发病情况

（2） 根据资料特点，计算每年龄组的患者人数及每 5 岁患者人数（见表 12-8），再绘制直方图。

表 12-14　224 例胸膜炎患者的年龄分布

年龄（岁）	患者人数	每 5 岁患者人数
11～	9	9
16～	30	30
21～	100	50
31～	61	30.5
41～	20	10
51～61	4	2
合　计	224	224

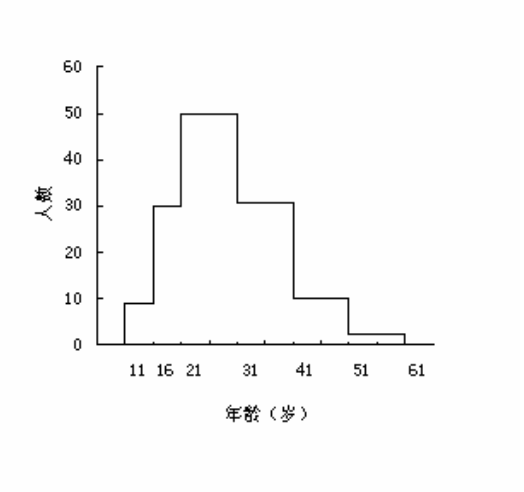


图 12-9　224 例胸膜炎患者年龄分布

4．根据题意，可列统计表 12-15 和统计图 12-10。

表 12-15　某县两年不同人群锡克试验反应结果分析

	1947 年			1975 年		
	调查人数	阳性人数	阳性率（%）	调查人数	阳性人数	阳性率（%）
幼儿园	144	37	25.69	101	21	20.79
小学生	1417	323	22.79	145	22	15.17
中学生	359	41	11.42	236	15	6.36

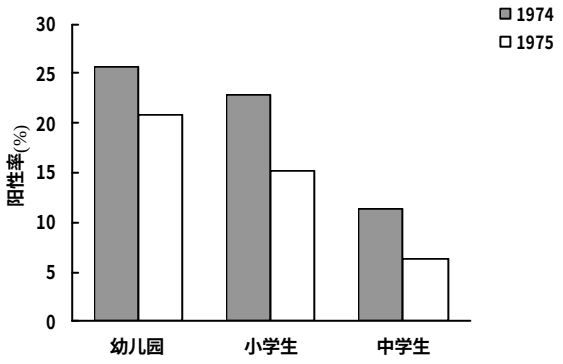


图12-10 某县两年不同人群锡克试验阳性率(%)

不同人群锡克试验反应阳性率均以 1974 年较高。

（四）判断正误并简述理由：

1. **正确。**

2．错。在一个统计表中，如果某处数据为“0”，就填“0”，如果数据暂缺则填“…”，若该处无数据，则填“—”。

3．正确。

4．正确。

5．错。 百分条图表示事物各组成部分在总体中所占的比重，以长条的全长为100%，然后按各构成比由大到小或由小到大排列绘图。

（颜艳　杨鹏）

第十三章　实验设计

一、教学大纲要求

（一）掌握内容

1．实验设计的基本原则

随机化原则、对照的原则（对照的类型,对照的设置）、重复的原则。

2．实验设计的基本内容和步骤

3．常用的实验设计方法

（1）随机化分组方法；

（2）完全随机分组设计；

（3）配对设计；

（4）配伍组设计及随机分组方法。

4． 确定样本含量

确定样本含量应当具备的条件： α、$1-\beta$、δ、σ p **。**

(二) 熟悉内容

1．常用的估计样本含量的计算方法及估计该试验的检验效能的方法。

（1）两样本均数比较。

（2）配对试验。

（3）样本均数与总体均数的比较。

（4）两样本率的比较。

（5）配对资料进行卡方检验时的样本含量估计。

（6）抽样调查估计总体均数的样本含量。

（7）抽样调查估计总体率的样本含量。

2．一致性检验：*Kappa* 值的意义及计算。

（三）了解内容

1 实验设计的特点和分类。

2.临床设计书的主要内容。

3.*Kappa* 值的抽样误差和假设检验。

二、教学内容精要

（一）实验设计的特点和分类

实验研究(experimental study)**是指研究者根据研究目的(或研究假设),主动加以干预措施，并观察总结其结果，回答假设研究所提出的问题的一种研究方法。**

实验研究可根据研究对象的不同分为两类：以动物或标本为研究对象的实验研究（experiment）**和以人为研究对象的临床试验**（clinical trial）**。**

（二）实验设计的基本原则

1.随机化原则

总体中的每一个观察单位都有同等的机会被选入实验组和对照组或进入样本，保证了非

处理因素在各组间均衡一致而使样本具有代表性。

2.对照原则

正确的设立对照可可控制实验过程中非实验因素的影响和偏倚，从而使处理因素的效应充分的显露出来。

设立对照组的常见方法有：空白对照、安慰剂（placebo）对照、实验对照、标准对照及自身对照。

3.重复的原则

保证每一个处理都有足够的重复数（样本量），避免把偶然性或巧合的现象当作必然的规律性现象，并能正确的估计实验误差。

（三）实验设计的基本内容和步骤

1.建立研究假设

在选题时应当考虑题目的科学性、新颖性、可行性以及所选课题是否是当前社会需要解决的主要问题。

根据研究目的确定本研究需要解决的主要问题（primary question）及相应的辅助问题（secondary question）。

2.明确研究范围

审慎考虑规定适当的纳入标准（inclusion criteria）和排除标准（exclusion criteria），选择适宜本次实验的受试对象。

3.确立处理因素

分清处理因素和非处理因素，并注意处理因素的标准化。

4.明确观察指标

选用客观性较强，易于量化，灵敏性和特异性均较高的指标。

5.控制误差和偏倚

采取各种有效措施控制误差（error）和偏倚（bias），使处理措施的效果能够真正的体现出来，是实验设计的重要任务之一。

（1）误差：泛指实测值与真值之差。

随机误差：随机误差（random error）它是一类不恒定的、随机、变化的误差，是不可避免的，但随机误差服从正态分布，可以用医学统计学的方法进行分析和推断。

系统误差：系统误差（systematic error）是指实验过程中产生的一些误差，它们的值是恒定不变或者是遵循着一定的规律变化。这两种误差都是人为因素产生的，可控制的。

（2）偏倚：属于系统误差，它是指在实验中由于某些非实验因素的干扰所形成的，歪曲了处理因素的真实效应。

选择性偏倚：选择性偏倚（selection bias）是由于纳入观察对象的方法不正确而产生的偏倚。它产生于实验研究的开始阶段，即研究对象的选择时产生。

测量性偏倚：测量性偏倚（measurement bias）是在实验过程中对研究对象进行观察或测量而造成的偏倚。它产生于实验进行的过程中。

在实验研究中，特别是在社区实验研究中，产生测量性偏倚的常见因素有：沾染（contamination）、干扰（intervention）、依从（compliance）和非依从（noncompliance）、失访（lost of follow-up）、检查和诊断结果的不一致（disagreement）、观察记录的失误、心理因素的干扰。防止测量性偏倚的主要方法：盲法（blind method）、签订实验合同、检查实验对

象的依从情况、注意医德、注意医德、定期检查研究记录、对每一种实验方法、诊断标准，重复判断的一致性应在实验前作出估计。

混杂性偏倚：混杂性偏倚（confounding bias）由于某些非实验因素与实验因素同时并存的作用影响到观察的结果，造成混杂性偏倚。它产生于总结分析阶段。

可通过对资料进行分层分析或采取配比法控制混杂性偏倚。

（四）常用的实验设计方法

1．随机对照试验

随机对照实验（randomized control trial）由于采取了随机化的分配原则，增强了各比较组间的可比性，避免了某些非实验因素的干扰，使实验因素能充分的显露出来，由于随机化，满足了统计学假设检验的要求，使检验结果更能反映它们之间存在的真实差异；设立对照组，更好的控制非实验因素对实验因素的影响，有效的控制了偏倚和误差，有利于反映所比较组间所存在的真实差异。

随机双盲对照实验，是目前国际上认为值得提倡的实验设计方案，特别适用于临床治疗效果、疾病的预后和诊断实验的研究。

2．配对设计

配对设计（paired design）可增强处理组间的均衡性，提高实验效率。

3．配伍组设计

又称随机单位区组设计（randomized block design）是配对设计的扩大（处理数大于2）。

（五）确定样本含量

1.确定样本含量的意义

确定适当的样本含量，可节约资源，并防止因为样本含量过少引起的检验效能偏低，出现了非真实的阴性结果，这是当前医学研究中值得注意的问题。

2.确定样本含量时应当具备的条件

建立检验假设；确立检验水准 α；提出在特定检验水准的条件下，所期望的检验效能 $1-\beta$；总体参数间的差值 δ；估计的总体标准差 σ及估计的总体率 p 。

3.确定样本含量的用途

保证科研设计有适当的样本含量，而且可考察当前的样本含量是否能够保证足够大的检验效能。

4.常用的估计样本含量的方法

（1）两样本均数比较

$$N=\left[\frac{2(u_a+u_b)s}{d}\right]^2 \qquad (13\text{-}1)$$

注意：上式中N为两组合计的样本含量，α有单双侧之分，β只取单侧。

$$u_b=\frac{d\sqrt{N}}{2s}-u_a \qquad (13\text{-}2)$$

上式是已知样本含量时（试验结束后），估计其检验效能是否足够大。方法是根据 u_b 查正态分布表得 b ，得到检验效能 $1-b$ 。

（2）配对试验

$$N=\left[\frac{(u_a+u_b)s_d}{d}\right]^2 \tag{13-3}$$

N 为观察的对子数。

估计检验效能：

$$u_b=d\frac{\sqrt{N}}{s_d}-u_a \tag{13-4}$$

（3）样本均数与总体均数的比较

$$N=\left[\frac{(u_a+u_b)s}{d}\right]^2 \tag{13-5}$$

$$u_b=d\frac{\sqrt{N}}{s}-u_a \tag{13-6}$$

（4）两样本率比较，当例数相等时

$$N=\frac{(u_a+u_b)^2 4p_c(1-p_c)}{(p_1-p_2)^2} \tag{13-7}$$

$$u_b=\frac{\sqrt{N}|p_1-p_2|}{2\sqrt{p_c(1-p_c)}}-u_a \tag{13-8}$$

式中p_1、p_2分别代表两组的总体率，p_c代表两组的合并率。N为两组合计之样本含量。

（5）配对分类资料多用 c^2 检验进行处理的资料的样本含量估计

$$N=\left[\frac{u_a\sqrt{2p_c}+u_b\sqrt{2p_{-+}p_{+-}/p_c}}{p_{-+}-p_{+-}}\right] \tag{13-9}$$

$$p_{+-}=\frac{b}{a+b}\qquad p_{-+}=\frac{c}{a+c}\qquad p_c=\frac{p_{+-}+p_{-+}}{2}$$

$$u_b=\frac{\sqrt{N}|p_{-+}-p_{+-}|-u_a\sqrt{2p_c}}{\sqrt{2p_{-+}p_{+-}/p_c}} \tag{13-10}$$

（6）抽样调查估计总体均数的样本含量

$$N=\left(\frac{u_a s}{d}\right)^2 \tag{13-11}$$

（7）抽样调查估计总体率的样本含量

$$N=\frac{u_a^2 p(1-p)}{d^2} \tag{13-12}$$

5.一致性检验 *Kappa* 值是判断一致性和信度评价的常用的重要指标。

$$Kappa=\frac{P_A-P_e}{1-P_e} \tag{13-13}$$

Kappa 值愈大，一致程度愈好，一般来说，Kappa 值≥0.75，说明已取得相当满意的一致程度，若小于 0.4，说明一致程度不够理想。

三、典型试题分析

（一）名词解释

实验效应。

答：实验效应（experimental effect**）主要指处理因素作用于实验对象的反应，这种效应将通过实验中观察指标显示出来。**

（二）填空题

实验研究与调查研究的区别在于　　。

答：前者主动施加干预措施而后者不。

[评析]实验研究是指研究者根据研究目的，主动加以干预措施，并观察总结其结果，回答研究假设所提出的问题。而调查研究旨在客观的描述总体，未加任何干预措施。

（三） 是非题

1. **实验效应选择特异性高的指标可减少假阳性率（　　）。**

答：正确。

[评析] 实验效应选择特异性高的指标减少假阳性率，而敏感度高的指标减少假阴性率。

2. **随机对照实验中所谓随机化就由受试对象随便选择进入实验组或对照组（　　）。**

答：错。

[评析] 随机不等于随便，所谓随机是指总体中的每一个观察单位都有同等的机会被选入样本或进入实验研究的各处理组中。

（四） 简答题

在选取实验效应时应考虑那几方面的问题？

答：应考虑选用客观性较强，易于量化，灵敏度高精确性较强的指标。

（五） 计算题

1. **为考虑某疫苗的疗效，拟进行一场实验，该传染病的发病率一般为** 10%**，接种组降低发病率** 5%**以上才有推广价值，问两组各需多少人？（**$a = 0.05$**,检验效能** 90%**）。**

答：由原题可知接种疫苗后只会降低发病率，a **=0.05（单侧），**$u_{0.05} = 1.64$**，** b **=0.10，**$u_{0.10}$**=1.28，**p_1 **=0.1，**$p_2 = 0.05$**，** d **=0.05，**p_c**=0.075**

代公式：

$$N = \frac{(u_a + u_b) \times 4p_c(1-p_c)}{(p_1 - p_2)} = \frac{(1.64+1.28) \times 4 \times 0.075 \times 0.925}{0.05^2} = 946$$ **人**

两组共需 946**，即每组** 473 **人。**

2. **新生儿的出生体重其均数为** 3200g**，标准差为** 467g**。欲研究妇女在怀孕期间服用某药物是否会影响新生儿体重，假设服用该药后出生的新生儿将比一般的新生儿平均增重** 220g**，假设单侧检验，**a **=0.05。问：**

（1） **如果取** 1- b **=0.08，两组样本含量相等时需要多大的样本含量才能发现其差异？**

（2） **如果** 1- b **为** 0.90**，取两组相等时，需要多大的样本含量？**

（3） **如果每组各有** 120 **人进入研究，仍采用单侧检验**a **=0.05，检验效能为多大？**

答：

（1）代入公式 $N=\left[\frac{2\left(m_a+m_b\right)s}{d}\right]^2=\left[\frac{2(1.64+0.84)467}{220}\right]^2$ =111

两组样本含量相等时，需要 112 例样本才能发现其差异。

（2）代入公式 $N=\left[\frac{2\left(m_a+m_b\right)s}{d}\right]^2=\left[\frac{2(1.64+1.28)467}{220}\right]^2$ =154

取两组相等时，需要 154 例样本。

（3）代入公式 $u_b=\frac{d\sqrt{N}}{2s}-u_a=\frac{220\sqrt{240}}{2\times467}-1.64=2.00$

解得：m_b =2.00，查表得：b =0.02，*power*=1-0.02=0.98。

3. 欲研究小剂量阿司匹林预防男性冠心病的效果，拟进行为期 5 年的随机双盲试验。若 40~64 岁男服用安慰剂后，冠心病 5 年发病率为 2.5%，同一年龄男性服用阿司匹林后冠心病 5 年发病率为 2.0%，问：

（1）取 a =0.05，用双侧检验，要有 80%的机会发现其差异，每组需要多少人进入研究？

（2）如检验效能取 0.90，其余的条件不变，各组又需多少人？

（3）若单侧检验，检验效能仍为为 0.80，各组又需多少人？

（4） 如每个组有 5000 人进入研究，a =0.05，双侧检验时期检验效能为多大？

答：

(1) a =0.05, u_a =1.96, b =0.20, u_b =0.84, p_1=0.025, p_2=0.02, p_c =0.0225

$$N=\frac{\left(u_a+u_b\right)^2 4p_c\left(1-p_c\right)}{\left(p_1-p_2\right)^2}=\frac{(1.96+0.84)^2 4\times0.0225\times(1-0.0225)}{(0.025-0.02)^2}=27588$$ 人

取 a =0.05，用双侧检验，要有 80%的机会发现其差异，每组需要 13794 进入研究。

（2） a =0.05, u_a =1.96, b =0.10, u_b =1.28

$$N=\frac{\left(u_a+u_b\right)^2 4p_c\left(1-p_c\right)}{\left(p_1-p_2\right)^2}=\frac{(1.96+1.28)^2 4\times0.0225\times(1-0.0225)}{(0.025-0.02)^2}=36942$$

如检验效能取 0.90，其余的条件不变，每组需 18471 少人。

（3） a =0.05, 单侧 u_a =1.64, b =0.20, u_b =0.84

$$N=\frac{\left(u_a+u_b\right)^2 4p_c\left(1-p_c\right)}{\left(p_1-p_2\right)^2}=\frac{(1.64+0.84)^2 4\times0.0225\times(1-0.0225)}{(0.025-0.02)^2}=21644$$

若单侧检验，检验效能仍为 0.80，各组需 10821 人。

（4） a =0.05, 双侧 u_a =1.96

$$u_b=\frac{\sqrt{N}\left|p_1-p_2\right|}{2\sqrt{p_c\left(1-p_c\right)}}-u_a=\frac{\sqrt{10000}\left|0.025-0.02\right|}{2\sqrt{0.0225(1-0.0225)}}-1.96=0.1686$$

查表得 b =0.4325，则双侧检验时其检验效能为 1-0.4325=0.5675。

4. 根据既往观察，人群接种某预防制剂后，体温高于 37.5℃的反应率为 10%，今欲推广使用，拟再次证实真实反映率是否为 10%，要求容许误差在真实反应率的 20%以内，a =0.05，b =0.10，问按单纯随机抽样需观察多少人？

答：

取 a =0.05（双侧），$u_{0.05}=1.96$，p_0=0.1，d =10%×20%=0.02，

$$N = p_0(1-p_0)(\frac{u_a}{d})^2 = 0.1\times(1-0.1)\frac{1.96^2}{0.02^2} = 864 \text{ 人}$$

需观察 864 人。

5．已知藏族中 HbsAg 阳性感染为 14.78%，现欲抽样检查了解拉萨地区藏族人的 HbsAg 阳感染率，要求误差不超过 1%，a =0.05, b =0.10，问需调查多少人？

答：

已知 a =0.05（双侧），$u_{0.05}=1.96$， p_0 =0.1478，d =0.01，代入公式可得：

$$N = p_0(1-p_0)\left(\frac{u_a}{d}\right)^2 = 0.1478(1-0.1478)\left(\frac{1.96}{0.01}\right)^2 = 4838 \text{ 人}$$

需调查 4838 人。

四、 习 题

（一） 名词解释

1. 安慰剂对照　2. 随机化　3. 混杂因素　4. 系统误差　5. 偏倚
6. 实验研究　7. 沾染　8. 干扰　9. 失访　10. 随机对照试验

（二） 填空题

1.实验设计的基本原则是_____，_____，_____。

2.决定样本含量的条件有_____，_____，_____，_____。

（三） 选择题

1.在下面各种实验设计中，在相同条件下最节约样本含量的是。（　）

A. 完全随机设计　　B. 配对设计
C. 配伍组设计　　D. 交叉设计

2.为研究新药“胃灵丹”治疗胃病（胃炎，胃溃疡）疗效，在某医院选择 50 例胃炎和胃溃疡病人，随机分成实验组和对照组，实验组服用胃灵丹治疗，对照组用公认有效的“胃苏冲剂”。这种对照在实验设计中称为（　）。

A. 实验对照　B. 空白对照　C. 安慰剂对照　D. **标准对照**

3.某医师研究丹参预防冠心病的作用，实验组用丹参，对照组用无任何作用的糖丸，这属于（　）。

A. 实验对照　B. 空白对照　C. **安慰剂对照**　D. 标准对照

4.某医师研究七叶一枝花治疗胃溃疡疗效时，实验组服用七叶一枝花与淀粉的合剂，对照组仅服用淀粉，这属于（　）。

A. **实验对照**　B. 空白对照　C. 安慰剂对照　D. 标准对照

5．实验设计的三个基本要素是（　）。

A. 受试对象、实验效应、观察指标　　B. 随机化、重复、设置对照
C. 齐同对比、均衡性、随机化　　D. **处理因素、受试对象、实验效应**

6．实验设计的基本原则（　）。

A. 随机化、盲法、设置对照　　B. 重复、随机化、配对
C. 随机化、盲法、配对　　D. **随机化、重复、设置对照**

7．实验设计和调查设计的根本区别是（　）。

A.实验设计以动物为对象　　B.调查设计以人为对象

C.实验设计可随机分组　　D.实验设计可人为设置处理因素

8．在（　）中，研究者可以人为设置各种处理因素；而在（　）中则不能人为设置处理因素。

A. 调查研究　　B. 社区干预试验

C. 临床试验　　D. 实验研究

（四） 是非题

1．用元参钩藤汤治疗 80 名高血压患者，服用半月后比服用前血压下降了 2.8kPa，故认为该药有效（　）。

2. 在实验设计中，样本含量越大，越符合其重复原则，越能降低实验误差（　）。

（五） 简答题

1．随机化的作用是什么？

2．某医师欲观察保健品“海兰兰”纠正小学生贫血的效果，您认为应采用何种类型的研究？在进行研究设计时应考虑那些主要问题，请简述之。

3．某单位研究饮食中缺乏维生素 E 与肝中维生素 A 含量的关系，将同种属的大白鼠按性别相同，年龄、体重相近者配成对子，共 8 对，并将每对中的两头动物随机分到正常饲料组和维生素 E 缺乏组，过一定时期将大白鼠杀死，测得其肝中维生素 A 的含量，问不同饲料的大白鼠肝中的维生素 A 的含量有无差别。请问：

（1）此实验属于那种实验设计（　）。

A.完全随机设计　　B.配对设计

C.配伍组设计　　D.拉丁方设计

（2）此实验结果应使用那种统计方法进行分析（　）。

A.配对资料 t 检验　　B. 回归分析

C. 成组资料 t 检验　　D. 成组设计方差分析

（3）以下假设检验那种是正确的（　）。

A. H_0 两种饲料喂养的大白鼠总体的肝中维生素 A 含量不等

H_1 两种饲料喂养的大白鼠总体的肝中维生素 A 含量相等

B. H_0 两种饲料喂养的大白鼠总体的肝中维生素 A 含量不等

H_1 两种饲料喂养的大白鼠总体的肝中维生素 A 含量相等

C. H_0 两种饲料喂养的大白鼠总体的肝中维生素 A 含量不等

H_1 两种饲料喂养的大白鼠总体的肝中维生素 A 含量不等

D. H_0 两种饲料喂养的大白鼠总体的肝中维生素 A 含量相等

H_1 两种饲料喂养的大白鼠总体的肝中维生素 A 含量不等

（4）结果如何解释（　）。

A. $P < 0.05$ 时，两组饲料喂养的大白鼠样本的肝中维生素 A 含量差别无意义

B. $P < 0.05$ 时，两组饲料喂养的大白鼠样本的肝中维生素 A 含量差别有意义

C. $P < 0.05$ 时，两组饲料喂养的大白鼠总体的的肝中维生素 A 含量差别无意义

D. $P < 0.05$ 时，两组饲料喂养的大白鼠总体的的肝中维生素 A 含量差别有意义

（六） 计算题

1.在进行有两种处理的动物冠状静脉窦的血流实验时，A 处理使平均血流量增加 1.8ml/min，B 处理使平均血流量增加 2.4ml/min。设两处理的标准差相等，均为 1.0ml/min，α=0.05，β=0.10，若要得出两处理有差别的结论，成组设计时需要多少实验动物？

2.据说某民族正常人平均体温高于 37℃，为核实这一点，拟进行抽样调查。如果就总体而言平均高出 0.1℃便不可忽略，已知正常人的体温标准差约为 0.2℃，那么，为了将第 I，II 类错误的概率 a 和 b 均控制在 0.05，试计算单纯随机抽样样本量应该是多大？

3.某药厂在大量筛选降压药物时规定平均降压效果超过 2kPa 者才作为候选药物进入下一轮研究。现对某药作了 10 个动物的预试验，血压下降值的标准差为 5kPa，问正式试验时样本量多大为宜？

4．为了比较两类片剂的溶解速率，决定各随机抽取 10 片，测定 5 分钟溶解量，然后作 a =0.05 水平的检验。据预试验，两类片剂的变异性相同，标准差约为 6 个单位，均数之差也约为 6 个单位，问欲使检验效能达到 95%，样本量应当多大？

5. 甲乙两医院的内科分别随机调查了 30 名住院病人，甲医院中对医疗服务表示满意者有 20 名，乙医院中表示满意者有 23 名。经统计检验，尚不能认为两医院内科住院病人的满意率不等。如欲考察两医院内科住院病人的满意率是否相差 10%以上，至少应当各调查多少病人？

6. 按 120 名患者就诊顺序，完全随机将其分为 A，B，C 三组。试列出随机分组表。试验结束后统计，发现其中有 56 个重症患者，就诊序号分别为：1～9，15～24，70～89，100～116。问 A，B，C 三组重症患者比例是否均衡？

五、习题答案要点

（一）名词解释

1.**安慰剂对照**（placebo control）指在实验研究中，对照组使用一种外形与实验药物完全相同而毫无药理作用的物质，这种对照称为安慰剂对照。

2.**随机化**（randomization）指研究对象中或总体中每一个观察单位都有同等的机会被选入样本或实验研究的各处理组中。

3.**混杂因素**（confounding factor）指实验研究中由于某些非实验因素与实验因素同时并存的作用影响到观察的结果，这种非实验因素称为混杂因素。

4.**系统误差**（systematic error）指实验过程中产生的一些误差，它们的值是恒定不变或者是遵循着一定的规律变化。

5.**偏倚**（bias）是指在实验中由于某些非实验因素的干扰所形成的系统误差，歪曲了处理因素的真实效应。

6.**实验研究**(experimental study)是指研究者根据研究目的（或研究假设），主动加以干预措施，并观察总结其结果，回答假设研究所提出的问题的一种研究方法。

7.**沾染**（contamination）是指对照组的实验对象接受实验组的处理措施，提高了对照组的有效率，其结果是造成了实验组和对照组之间差异缩小的假象。

8.**干扰**（intervention）是实验组从实验外接受了对实验因素有效的药物或措施（非处理措施），提高了实验组的有效率，其结果是扩大了实验组和对照组之间的差异。

9．失访（lost of follow-up）指受试者在实验过程中由于各种原因退出实验称为失访。

10.随机对照实验（randomized control trial）首先将受试对象随机分配到实验组和对照组，通过比较分析回答研究假设的问题。

（二）填空题

1.重复、对照、随机化。

2.检验水准 α、检验效能 $1-\beta$、总体参数间的差值 δ、估计的总体标准差 σ。

（三）单项选择题

1. D　2. D　3. C　4. A　5. D　6. D　7. D　8. BC, A

（四）是非题

1．错。没有设立对照不能说明问题。

2．错。样本含量过大，实验过程不易控制，反而增加系统误差，且成经济损失，故样本含量适当时，效能最高，重复性原则并非指样本含量越大约好。

（五）简答题

1．随机化保证了各比较组间的均衡可比性，也是资料统计分析时进行统计推断的前提。

2．宜采用配对设计，将实验对象按照年龄，性别，营养状况，贫血轻、中、重的程度配对，随机分配每对中两个对象接受不同的处理方式。实验组给予“海兰兰”对照组给予安慰剂，最好采用双盲法。

3．（1）B　（2）A　（3）D　（4）D

（六）计算题

1.本题 δ=2.4-1.8=0.6ml/min，σ=1，双侧 α=0.05，β=0.1。查表得 $u_{0.05}=1.96$，$u_{0.01}=1.282$，按两组均数 t 检验估计样本含量：

$$N=\left[\frac{2(u_a+u_b)s}{d}\right]^2=\left[\frac{2(1.96+1.282)\times 1}{0.6}\right]^2=120$$ 只

共需 120 只，每组 60 只。

2．由原题可知 a=0.05（单侧），$u_{0.05}=1.64$，b=0.05，$u_{0.05}=1.64$，d=0.1 ^{0}C，s=0.2 ^{0}C，按样本均数与总体均数比较 t 检验估计样本含量：

$$N=\left[\frac{(u_a+u_b)s}{d}\right]^2=\left[\frac{(1.64+1.64)\times 0.2}{0.1}\right]^2=43$$

可取 43 个人参加试验。

3.由原题可取 a=0.05（单侧），$u_{0.05}=1.64$，b=0.01，$u_{0.01}=2.33$，d=2kPa，s=5kPa，因为 s 未知，所以用 S 代替 σ，按配对 t 检验估计样本含量：

$$N=\left[\frac{(u_a+u_b)s}{d}\right]^2=\left[\frac{(1.64+2.33)\times 5}{2}\right]^2=98$$

可取 98 只动物。

4.由原题可知 a=0.05（双侧），$u_{0.05}=1.96$，b=1-0.95=0.05，$u_{0.05}=1.64$，d=6，s=6，按两组均数 t 检验估计样本含量：

$$N=\left[\frac{2(u_a+u_b)s}{d}\right]^2=\left[\frac{2\times(1.96+1.64)\times 6}{6}\right]^2=51$$

总片数只需 52 片，每类 26 片。

5.由原题可取α **=0.05（双侧），**$u_{0.05}=1.96$，β **=0.10，** $u_{0.10}=1.28$，$p_1=\frac{2}{3}$，$p_2=\frac{23}{30}$，$p_c=\frac{43}{60}$，**代入公式得：**

$$N=\frac{(u_\alpha+u_\beta)^2 4p_c(1-p_c)}{(p_1-p_2)^2}=\frac{(1.96+1.28)^2\times 4\times\frac{43}{60}\times(1-\frac{43}{60})}{(\frac{2}{3}-\frac{23}{30})^2}=852$$

每个医院各调查 426 **人。**

6.用计算器给每个患者产生一个 3 位数的随机数，规定随机数区间，000～332 分到 A 组，333～665 分到 B 组，666～998 分到 C 组。随机分组表见表 13-1。

表 13-1　120 **例患者随机分组结果**

患者编号	随机数字	分组结果	患者编号	随机数字	分组结果	患者编号	随机数字	分组结果	患者编号	随机数字	分组结果
1*	628	B	31	747	C	61	647	B	91	994	C
2*	673	C	32	791	C	62	474	B	92	507	B
3*	833	C	33	503	B	63	685	C	93	542	B
4*	915	C	34	568	B	64	414	B	94	309	A
5*	776	C	35	442	B	65	878	C	95	871	C
6*	713	C	36	002	A	66	790	C	96	375	B
7*	366	B	37	735	C	67	201	A	97	701	C
8*	663	B	38	598	B	68	690	C	98	141	A
9*	830	C	39	400	B	69	703	C	99	305	A
10	842	C	40	157	A	70*	723	C	100*	018	A
11	123	A	41	531	B	71*	437	B	101*	341	B
12	318	A	42	820	C	72*	126	A	102*	769	C
13	168	A	43	801	C	73*	222	A	103*	334	B
14	461	B	44	125	A	74*	010	A	104*	125	A
15*	449	B	45	503	B	75*	109	A	105*	292	A
16*	658	B	46	692	C	76*	479	B	106*	314	A
17*	123	A	47	112	A	77*	648	B	107*	957	C
18*	532	B	48	370	B	78*	947	C	108*	322	A
19*	993	C	49	443	B	79*	875	C	109*	842	C
20*	661	B	50	465	B	80*	120	A	110*	445	B
21*	394	B	51	911	C	81*	236	A	111*	412	B
22*	571	B	52	601	B	82*	873	C	112*	874	C
23*	931	C	53	265	A	83*	010	A	113*	523	B
24*	174	A	54	520	B	84*	923	C	114*	499	B
25	785	C	55	502	B	85*	391	B	115*	421	B
26	329	A	56	129	A	86*	436	B	116*	748	C
27	321	A	57	484	B	87*	786	C	117	945	C
28	700	C	58	560	B	88*	562	B	118	797	C
29	443	B	59	294	A	89*	919	C	119	485	B
30	690	C	60	948	C	90	536	B	120	508	B

备注：*为重症患者。

从表 13-1 中可统计出 A 组、B 组、C 组中重症患者数分别为 14、22、20 个，A、B、C 三组重症患者分布的均衡性检验结果见表 13-2。

13-2　A、B、C 三组重症患者分布的均衡性检验

分组	人数		合计
	重症患者	轻度患者	
A 组	14	16	30
B 组	22	26	48
C 组	20	22	42
合计	56	64	120

计算得 c^2 =0.028，$c^2_{0.05,2}$ =5.99，不能认为 A、B、C 三组重症患者分布不均衡。

（周燕荣　陈平）

第十四章　调查设计与资料分析

一、教学大纲要求

(一) 掌握内容

1． 调查的概念及其特点，调查研究与实验研究的区别。

2． 调查设计的基本原则与内容

(1) 明确调查目的。

(2) 确定调查对象和观察单位。

(3) 确定调查方法。

(4) 确定调查指标和变量。

(5) 调查工具和调查表的种类、调查表和问卷的一般结构、调查问题的形式、调查问题设计应注意的问题。

(6) 确定样本含量的意义及方法。

(7) 对调查员的要求。

(8) 有关伦理道德的问题。

3． 常用的抽样方法

(1) 概率抽样的概念。

(2) 常用的概率抽样方法：简单随机抽样、系统抽样、分层抽样、整群抽样。

(3)

(二)　熟悉内容

(1) 非概率抽样的概念，配额抽样、“滚雪球”样本、识别(判断)样本的概念。

(2) 标准化率的概念及计算方法。

(三)　了解内容

病例对照研究和队列研究的概念及数据的处理和分析。

二、教学内容精要

(一) 调查的概念及其特点

调查(survey)是指在没有任何干预措施的条件下客观地观察和记录研究对象的现状及其相关特征。

在调查中，欲研究的对象及其相关特征(包括研究因素和非研究因素)是客观存在的，不能采用随机分配的方法来平衡或消除非研究因素对研究结果的影响，这是调查研究区别于实验研究的最重要特征。当然对非研究因素的控制可以在调查分析阶段通过标准化法、分层分析以及多因素统计分析等方法得以实现，而不是在调查阶段。

(二) 调查设计的基本原则与内容

1．明确调查目的

每一项调查，必须有明确的调查目的。调查目的一般可分为调查的总目的和具体目的。调查

目的是选定调查指标的依据。

2．确定调查对象和观察单位

根据调查目的确定调查对象，即明确调查总体的同质范围。在确定的总体范围内，组成调查对象的每个个体即为观察单位。观察单位可以是一个人、一个家庭或一个群体。

3．确定调查方法

根据研究问题的性质、客观条件和研究目的选择合适的调查方法。按调查的涉及面，一般可分为普查(overall survey)和抽样调查(sampling survey)。普查也称全面调查(complete survey)，是对调查范围内的全部观察对象(总体)进行调查，一般用于了解总体在某一特定“时点”的情况。抽样调查是一种非全面调查，是从总体中抽取一定数量的观察单位组成样本，然后根据样本信息来推断总体特征。抽样调查是医学科研中最为常用的方法。

调查方法还可按调查的内容发生的时间，分为横断面调查(cross-sectional study)和纵向调查(longitudinal study)；按资料的来源，可分为现场调查和利用现有资料两种；按调查方式，可分为面对面调查和非面对面调查(信访、电话采访等)两种。

4．确定调查指标和变量

调查目的是选定调查指标的依据，调查指标是调查目的的具体体现。设计时，应将调查目的转化为具体的调查指标。调查指标可分为客观指标和主观指标，还可分为定性指标和定量指标。一个指标可以是一个或几个变量，也可以是几个指标构成一个变量。指标的设立应注意灵敏性、特异性和客观性，并紧扣研究目的，做到少而精。

5．调查工具和调查表

（1）调查工具：调查工具(instruments)可分为两类，一类是“硬”工具，一类是“软”工具。如尺、秤、温度计等是“硬”工具；调查表和问卷等是“软”工具。调查工具必须标准化，要防止系统误差。工具的使用和调查结果的记录也必须标准化。

（2）调查表和问卷的一般结构：调查表和问卷(questionnaire) 一般可划分为 4 个部分，分别为：“说明部分”、“填写说明”、“核查项目”、“调查项目”。“说明部分”主要说明调查目的，以取得调查对象的合作；“填写说明”是为了保证所有调查员和调查对象均能对调查项目及填写方法正确理解、统一认识而编写的；“核查项目”这一部分是与调查目的无关、不向调查对象询问的质量控制项目，如调查员姓名、调查日期、复核结果、未调查原因等；“调查项目”部分是调查对象填写的部分，是调查的核心内容。

（3）问题的形式：根据问题答案的形式，问题可分为开放型和封闭型两类。开放型问题对问题答案不加任何限制，由调查对象对问题自由回答，适于调查者不清楚答案如何以及答案很多的情况，或事先不能确定回答的范围以及预调查；封闭型问题是根据问题可能的答案，提出两个或多个固定答案共调查对象选填，常用“是与否”或多项选择的形式。封闭型问题只能得到分类资料或等级资料，而开放型问题有时可得到数值变量资料。可根据具体情况加以选择。

（4）问题设计应注意的问题：尽量避免术语；避免含糊的提问用词；避免双重问题；避免诱导或强制；敏感问题的调查要有专门技巧。

6．确定样本含量

为什么要确定样本含量或者说其意义有哪些？

(1) 可以控制统计量的抽样误差，样本含量越大，标准误越小；

(2) 提高估计的精度，增大样本含量是控制可信区间的宽度的有效办法；

(3) 增大样本含量是控制统计分析中II型错误的概率大小的有效措施；

(4) 表示抽样误差的指标（各种标准误）的抽样误差也与样本含量有关（如样本方差的方差)。

在现场调查中，最常用的是估计总体均数及估计总体率时要求的样本含量。

估计总体均数的样本含量的计算公式：

$$n=(\frac{t_{a/2}s}{d})^2 \qquad (14\text{-}1)$$

式中d 为允许误差。s 为估计的标准差，一般都是从以前的研究资料中获得。在算得n 之前，自由度n 不能确定，$t_{0.05/2}$ 仍是未知的，解决的办法是先以$u_{0.05/2}$ 代替 $t_{0.05/2}$，用迭代法求得n。

估计总体率的样本含量的计算公式：

$$n=\frac{u^2_{a/2}p(1-p)}{d^2} \qquad (14\text{-}2)$$

式中d 为允许误差。如果估计的 p 是一个范围，那就应该取其中最靠近 50%的值。假定估计的 p 约在 10%到 30%之间，则取 p =0.30；假定估计的 p 约在 40%到 80%之间，则取 p =0.50；如果对 p 一无所知，则取 p =0.50。

7．调查员

调查员应该经过选择和培训，培训分理论培训和实践培训。调查员的工作量要合理，对调查员应有监督机制和质量控制措施。

8．伦理道德

伦理道德问题可以来自于某些调查的问题本身，也可以来自于为获得有效而可靠的资料所用的方法。调查时要注意知情同意（informed consent）和隐私的保护。知情同意是指在研究对象暴露于某种危险之中或丧失某种个人权益时，要征得研究对象同意。

(三）抽样方法

1．概率抽样

所谓概率抽样(probability sampling)，就是在抽样中必须使该总体中的每一个个体都有已知的或可计算的和非零的概率被抽样抽中。常用的概率抽样方法包括：简单随机抽样、系统抽样、分层抽样和整群抽样。各种抽样方法的抽样误差一般是：整群抽样≥简单随机抽样≥系统抽样≥分层抽样。在应用大多数的概率抽样方法时，确切的抽样框架非常重要。抽样框架(Sampling frame)，简单地说就是一份完整的可以用来抽样的名单。如果没有抽样框架，也就是说目标人群（总体）不明确，那么所得的调查结论很难说适用于什么人群。

（1）简单随机抽样：所谓简单随机抽样(simple random sampling)是在某个总体中以完全随机的方法抽取一部分个体组成样本。一般，在抽样前，需要先对抽样总体中的全部个体进行编号即确定抽样框架，然后用抽签或随机数字表的方法抽取一部分个体。这种抽样方法简单，计算抽样误差方便。但是，在大规模的调查中，由于对总体中的所有个体进行编号很困难，而且当样本量不大时抽取的个体可能很分散，因此，抽样和现场调查都会相当困难。

简单随机抽样的均数和率的标准误的计算公式如下：

$$S_{\bar{X}} = \sqrt{(1-\frac{n}{N})\frac{S^2}{n}} \qquad (14\text{-}3)$$

$$S_p = \sqrt{(1-\frac{n}{N})\frac{p(1-p)}{n-1}} \qquad (14\text{-}4)$$

期中，n/N 称为抽样比(sampling fraction)，$(1-n/N)$ 为“有限总体校正数”(finite population correction)。去掉“有限总体校正数”即可用于无限总体抽样误差的计算。

（2）系统抽样：所谓系统抽样(systematic sampling)是指随机地在抽样框架内每间隔若干个个体抽取一个个体的抽样方法。在一般情况下，系统抽样的抽样误差是和简单随机抽样相仿甚至比简单随机抽样的抽样误差更小。系统抽样的抽样误差一般按简单随机抽样方法估计。

（3）分层抽样：所谓分层抽样(stratified sampling)是先按对观察指标影响较大的某种特征，将总体分为若干类别(统计上称之为“层”，strata)，再从每一层内随机抽取一定数量的观察单位，合起来组成样本。分层的原则是层间差别越大越好，层内差别越小越好。在样本总含量 n 确定后，有两种比较常用的方法来分配各层的观察单位数 n_i 。

按比例分配(proportional allocation)：按各层观察单位数 N_i 占总体观察单位数 N 比例抽取样本，使各层样本含量 n_i 与样本总含量 n 之比等于各层观察单位数 N_i 与总体观察单位数 N 之比。采用按比例分层随机抽样时，所得均数或比例是自动加权的。样本量分配可按下式计算：

$$\frac{n_i}{n} = \frac{N_i}{N} \quad \text{或} \quad n_i = N_i\frac{n}{N} \qquad (14\text{-}5)$$

最优分配(optimum allocation)：即同时按总体各层观察单位数 N_i 的多少和标准差 σ_i 的大小分配，按下面两式分配各层的样本量，使抽样误差最小。

均数的抽样公式：

$$n_i = n\frac{N_i s_i}{\sum N_i s_i} \qquad (14\text{-}6)$$

率的抽样公式：

$$n_i = n\frac{N_i\sqrt{p_i(1-p_i)}}{\sum N_i\sqrt{p_i(1-p_i)}} \qquad (14\text{-}7)$$

分层抽样中，若令 $W_i = N_i/N$，样本均数 $\bar{X}$ 和率 p 及其标准误的计算公式如下：

样本均数： $$\bar{X} = \sum W_i\bar{X}_i \qquad (14\text{-}8)$$

样本均数的标准误： $$S_{\bar{X}} = \sqrt{\sum(1-\frac{n_i}{N_i})W_i^2 S_{\bar{X}_i}^2} \qquad (14\text{-}9)$$

样本率： $$p = \sum W_i p_i \qquad (14\text{-}10)$$

样本率的标准误： $$S_p = \sqrt{\sum(1-\frac{n_i}{N_i})W_i^2 S_{p_i}^2} \qquad (14\text{-}11)$$

（4）整群抽样：所谓整群抽样(cluster sampling)是先将总体按照某种与研究指标无关的特征化分为 K 个群组，每个群包括若干观察单位，然后在随机抽取 k 个群，将抽取的各个群的全部观察单位组成样本。整群抽样的特点是抽样和调查都很方便，可能省时、省力和省钱。缺点是可能抽样误差较大，特别是群间差别较大时。

整群抽样样本均数 $\overline{X}$ 和率 p 及其标准误的计算公式如下：

样本均数：
$$\overline{X} = \frac{K}{Nk}\sum m_i \overline{X}_i \qquad (14\text{-}12)$$

均数的标准误：
$$S_{\overline{X}} = \frac{K}{N}\sqrt{(1-\frac{k}{K})(\frac{1}{k(k-1)})\sum_{i=1}^{k}(T_i-\overline{T})^2} \qquad (14\text{-}13)$$

式中 T_i 为样本第 i 群内观察值之和，$\overline{T}$ 为各 T_i 的均数，$\overline{T} = \sum T_i / k$。

样本率：
$$p = \frac{K}{Nk}\sum a_i \qquad (14\text{-}14)$$

率的标准误：
$$S_p = \frac{K}{N}\sqrt{(1-\frac{k}{K})(\frac{1}{k(k-1)})\sum_{i=1}^{k}(a_i-\overline{a})^2} \qquad (14\text{-}15)$$

式中 $\sum a_i$ 为样本中各群阳性数之和，$\overline{a}$ 为样本各群的平均阳性数。

2．非概率抽样

所谓非概率抽样(non-probability sampling)，是指各个个体被抽样抽中的概率是未知的和无法计算的。然而，一些非概率抽样方法，尽管不能按常规的理论来计算抽样误差和推断总体，在特定条件下，还是有用的。但在应用中，不能忘了它们的局限性，特别要注意结论的合适性。

（1）配额抽样：所谓配额抽样(quota sampling)是一种的实用的非概率抽样方法。就是要求样本中个体的构成在指定的几个特征方面的（分配额度）比例完全与总体一样，例如，由于全人口中男女各半，所以要求调查对象中也是男女各半，由于该地有苗族居民 30%，要求在调查对象中苗族居民占 30%。配额抽样可以使样本有宏观上的代表性。

(2)“滚雪球”样本和识别（判断）样本：在有些情况下，缺少目标总体中全部个体的名单，无法构成抽样框架，此时可用另外一些非概率抽样的方法，即“滚雪球”(snowballing)抽样和识别(judgement)抽样的方法。比如调查太极拳爱好者，由于正式参加太极拳运动的人数太少，因此难以获得抽样框架。但是每一位太极拳运动爱好者都会有一些相同兴趣的好友，所以可以通过这种关系滚雪球似地把样本扩大。所谓识别抽样，是指研究者尽可能找到和识别需调查的个体。这两种调查方法，都未能明确规定抽样框架，甚至难以说出要推断的总体是什么，然而，作为一项探索性的调查，仍可能获得有价值的信息。

(四) 病例对照研究和队列研究

病例对照研究(case control study)是一种“由果推因”的回顾性观察性研究，根据有无研究疾病或其它结局，将研究人群分为病例组(cases)和对照组(controls)，追溯过去某些暴露情况，比较两组暴露水平有无差异，从而得出结局与暴露有无关联的推断。

队列研究(cohort study)是一种“由因寻果”的纵向前瞻性观察研究。根据观察开始时有无暴露(exposure)史，研究者将没有出现研究疾病或其它结局(outcome)的研究人群分为暴露人群和非暴露人群，并随访观察一定时期，旨在比较两组人群的疾病“发病”率有无差异，从而得出暴露与结局有无关联的推断。两者关系可简要见下表：

表 14-1　病例对照研究与队列研究的比较

比较项目	病例对照研究	队列研究
观察方向	“由果推因”的回顾性观察研究	“由因寻果”的纵向前瞻性观察研究
可获得指标	比数比(odds ratio，OR)，当发病率很低时，OR 被认为与 RR 很接近； 用 OR 替代 RR 估计归因危险度百分比(attributable risk proportion，AR%)	累积发病率(cumulative incidence，CI)； 发病密度(incidence density，ID)； 相对危险度(relative risk，RR)； 归因危险度(attributable risk，AR)； 归因危险度百分比(attributable risk proportion，AR%)
优点	省时、省人力、省经费，易组织实施； 适于结局为罕见事件的病因研究； 一次调查可探索疾病的多个可疑病因，常用于初步验证某病因假说或探测某些病因； 当发病率很低时，OR 与 RR 相当近似	因结局发生在后，故对暴露资料的收集是无偏倚的； 可收集已知混杂因素的信息； 可直接计算发病率、相对危险度等疾病与病因关联的指标； 病因在前结果在后，可证实病因假说； 可获得多种结局资料
缺点	不适于研究人群中暴露比例很低的因素； 不能直接计算发病率； 有时难以判断暴露与疾病之间的时间先后关系； 易发生选择偏倚、回忆偏倚、混杂偏倚	耗费时间、人力、经费； 当结局为罕见事件时，需样本量大； 易产生以下偏倚：研究对象依从性偏倚、信息偏倚、对暴露与结局的评价偏倚

(五)标准化率

调查资料在进行对比分析时，要注意组间的可比性。当两组(或多组)资料的内部各小组的率明显不同，而且各小组观察单位的构成比明显不同时，则不能直接比较两组的总率。这时可采用一个“统一的标准”将两组(或多组)资料的内部构成比例调整一致后，分别计算出调整后的总率再作比较，这种方法叫做率的标准化(standardization)。率的标准化有以下两种方法：

1．直接标准化

直接标准化(direct standardization)是以有代表性的、人群数量大的组作为标准人群，用标准人群各小组观察单位数分别乘以被标化人群的各小组的阳性率(如发病率)，得到被标化人群的理论阳性数。理论阳性数除以标准人群总人口数，得到被标化人群的标准化阳性率。

2．间接标准化

间接标准化(indirect standardization)是以标准人群各小组阳性率乘以被标化人群的各小组观察单位数，得到被标化人群的理论阳性人数。被标化人群的实际阳性人数除以理论阳性总人数，得标准化阳性率比值(如标准化发病率比值standardized incidence ratio，SIR 或标准化死亡率比值 standardized mortality ratio，SMR)。SIR(SMR)乘以标准人群实际阳性率，得到

被标化人群的间接标准化阳性率。

3．计算符号及公式

表 14-2　计算用数据符号

组别	被标化组			标准组		
	观察单位数	阳性数	率	观察单位数	阳性数	率
1	n_1	r_1	p_1	N_1	R_1	P_1
2	n_2	r_2	p_2	N_2	R_2	P_2
3	n_3	r_3	p_3	N_3	R_3	P_3
⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	n_i	r_i	p_i	N_i	R_i	P_i
⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	n_k	r_k	p_k	N_k	R_k	P_k
合计	n	r	p	N	R	P

直接法计算公式：

$$p' = \frac{\sum N_i p_i}{N} \tag{14-16}$$

间接法计算公式：

$$p' = P\frac{r}{\sum n_i P_i} \tag{14-17}$$

$$SMR = \frac{r}{\sum n_i P_i} \tag{14-18}$$

三、典型试题分析

（一）名词解释

调查。

答案：调查(survey)是指在没有任何干预措施的条件下客观地观察和记录研究对象的现状及其相关特征。在调查中，欲研究的对象及其相关特征(包括研究因素和非研究因素)是客观存在的，不能采用随机分配的方法来平衡或消除非研究因素对研究结果的影响，这是调查研究区别于实验研究的最重要特征。

[评析]　本题考点：调查的概念及其特点，调查研究与实验研究的区别。

实验研究可以人为地设置干预措施，而调查研究是在没有任何干预措施的条件下观察和记录研究对象的现状及其相关特征。

（二）单项选择题

1．随机抽样是指（　　）。

A．每个个体必须有同样的概率被抽样抽中

B．抽样中不要按主观意愿挑选

C．概率抽样和非概率抽样

D．哪一个个体被抽样抽中完全是由于碰巧

答案：A

[评析] 本题考点：统计学中随机抽样概念的理解。

有限总体在获得抽样框架后，可以实现随机抽样，即总体中的每个观察单位都有同样的机会被选作样本，而不是按主观意愿挑选或哪一个个体被抽样抽中完全是由于碰巧。

2．概率抽样是指（ ）。

A． 每个个体被抽样抽中的概率是已知非零的

B． 每个个体被抽样抽中的概率是可计算的

C． 每个个体被抽样抽中的概率是非零的

D． 每个个体被抽样抽中的概率是非零的，已知或可计算的

答案：D

[评析] 本题考点：统计学中概率抽样概念的理解。

概率抽样就是在抽样中必须使该总体中的每一个个体都有已知的或可计算的和非零的概率被抽样抽中。常用的概率抽样方法包括：简单随机抽样、系统抽样、分层抽样和整群抽样。在应用大多数的概率抽样方法时，确切的抽样框架非常重要。概率抽样可以估计其抽样误差的大小。 而非概率抽样是指各个个体被抽样抽中的概率是未知的和无法计算的。

3．在常用的几种抽样调查中，其抽样误差的大小关系一般是（ ）。

A．整群抽样小于分层抽样

B．系统抽样大于简单随机抽样

C．整群抽样大于简单随机抽样

D．简单随机抽样小于最优分配分层随机抽样

答案：C

[评析] 本题考点：常用的几种随机抽样调查方法的抽样误差的估计。

常用的几种随机抽样调查方法有统计的理论依据，可估计抽样误差，能客观地评价调查结果的精度。各种抽样方法的抽样误差一般是：整群抽样≥ 简单随机抽样≥ 系统抽样≥ 分层抽样。在保证同样精度的条件下，所用抽样方法的抽样误差越大，则所需样本含量相对越多。

4．最优分配分层抽样，（ ）。

A．可以使抽样误差最小　　B．可以使调查费用最小

C．样本均数是无偏的　　D．要求的样本含量最小

答案：A

[评析] 本题考点：分层抽样调查抽样误差的估计。

分层抽样中，在样本总含量确定后，有两种比较常用的方法来分配各层的观察单位数。一种是按比例分配，另外一种是最优分配。按比例分配是按各层观察单位数占总体观察单位数比例抽取样本，使各层样本含量与样本总含量之比等于各层观察单位数与总体观察单位数之比；而最优分配是同时按总体各层观察单位数的多少和标准差的大小分配，使抽样误差最小。

5．调查设计和实验设计的根本区别是（ ）。

A．实验设计以动物为对象

B．调查设计以人为对象

C．实验设计可随机分组

D．实验设计可人为设置处理因素

答案：D

[评析] 本题考点：调查研究和实验研究的概念的理解。

调查是指在没有任何干预措施的条件下客观地观察和记录研究对象的现状及其相关特征。在调查中，欲研究的对象及其相关特征是客观存在的，不能采用随机分配的方法来平衡或消除非研究因素对研究结果的影响，而实验研究可人为设置处理因素，这是调查研究区别于实验研究的最重要特征。

（三）简答题

四种基本抽样方法是如何体现随机性的？各自的优缺点和适用的场合是什么？

答案：简单随机抽样：将调查总体的全部观察单位编号，再用随机数字标或抽签等方法随机抽取部分观察单位组成样本。优点：均数或率及标准误的计算简便。缺点：总体例数较多时，一一编号比较麻烦，实际工作中难以办到。适用场合：一些比较单纯的现象，如观察单位在总体中分布比较均匀时采用这种方法。

系统抽样：将总体的观察单位按某一顺序等分成n个部分，在从第一部分随机抽第k号观察单位，依次用相等间隔机械地从每一部分各抽一个观察单位组成样本。优点：易于理解，简便易行；容易得到一个按比例分配的样本。缺点：当总体的观察单位按顺序有周期趋势或单调增(减)趋势,则系统抽样产生明显偏性；没有自己的估计抽样误差的方法。适用场合：观察单位分布十分均匀，可以保证样本对总体有较好的代表性。

整群抽样：先将总体划分为K个群，每个群包含若干观察单位，再随机抽取k个群，并将被抽取的各个群的全部观察单位组成样本。优点：便于组织，节省经费。缺点：例数一定时，抽样误差较大。适用：群间差异较小的对象。

分层抽样：按影响观察值变异较大的某种特征，将总体化分为若干类型或组别(即层)，再从每一层内随机抽取一定数量的观察单位，合起来组成样本。优点：减少抽样误差；便于对不同的层采用不同的抽样方法；可以对不同层独立进行分析。适用：各层间差异较大。

[评析] 本题考点：常用的几种概率抽样调查方法的比较。

根据研究问题的性质、客观条件和研究目的选择合适的抽样调查方法。不同的抽样方法有不同的使用场合。

（四）计算题

某医师打算研究正常女大学生的收缩期血压(kPa)，要求本次调查所得样本均数与未知的总体均数相差不大于 0.5 的概率是 95%，以前的调查资料显示，标准差在 2.2(kPa)左右，若作简单随机抽样，需调查多少对象？

答案：本题为调查总体均数的样本含量估计。已知：

s =2.2，d =0.5，双侧$u_{0.05}$=1.96

根据公式$n=(\frac{t_{a/2}s}{d})^2$计算样本含量，但在算得n之前，自由度n不能确定，$t_{0.05/2}$仍是未知的，解决的办法是先以$u_{0.05/2}$代替$t_{0.05/2}$，用迭代法求得n。

首先根据公式$n=(\frac{u_{a/2}s}{d})^2$，n=（1.96×2.2/0.5）^{2}=74.37≈74（人）

由n=74，得自由度n=74-1=73，用n=73 查t界值表得$t_{0.05/2,73}$=1.993，再依据公式

$$n=(\frac{t_{a/2}s}{d})^2$$=（1.993×2.2/0.5）²=76.90≈77（人）

由n=76，得自由度n=76-1=75，用n=75查t界值表得$t_{0.05/2,75}$=1.992，再依据公式

$$n=(\frac{t_{a/2}s}{d})^2$$=（1.992×2.2/0.5）²=76.82≈77（人）

因此认为，调查样本含量为77人。

[评析] 本题考点：调查总体均数的样本含量估计。

在估计调查总体均数的样本含量时可用公式$n=(\frac{t_{a/2}s}{d})^2$采用迭代法求得，当然也可直接利用公式$n=(\frac{u_{a/2}s}{d})^2$求得。

四、习题

（六） 名词解释

1．抽样调查　2．简单随机抽样　3．系统抽样　4．分层抽样
5．整群抽样　6．概率抽样　7．非概率抽样　8．相对危险度
9．病例对照研究　10．队列研究

（二） 单项选择题

1． 在抽样调查中，理论上样本含量大小与（　）大小有关。

A．样本极差　B．样本变异系数
C．样本方差　D．样本四分位间距

2． 在计算简单随机抽样中估计总体均数所需样本例数n时，至少需要确定（　）。

A． 允许误差d，总体标准差s，第二类错误b
B． 第一类错误a，总体标准差s，总体均数m
C． 允许误差d，总体标准差s，第一类错误a
D． 允许误差d，总体标准差s，总体均数m

3． 拟用放射免疫法检测某人群(5000人)血液中流脑特异免疫球蛋白含量，根据文献报道，其标准差约为0.5mg/L，容许误差为0.1mg/L，则按单纯随机抽样，需抽出的样本例数为（　）人。

A．97　B．95　C．96　D．94

4．在抽样调查中，理论上样本含量大小会影响（　）。

A．样本标准差的大小　B．总体均数的稳定性
C．样本标准差的稳定性　D．样本中位数的大小

5．$S_{\bar{x}}=S/\sqrt{n}$表示（　）抽样时均数的抽样误差。

A．整群　B．系统　C．分层　D．简单随机

6．我们工作中常采用的几种抽样方法中，最基本的方法为（　）；

7．操作起来最方便的为（　）；

8．在相同条件下抽样误差最大的为（　）；

9．所得到的样本量最小的为（　）。

A．简单随机抽样　　B．系统抽样

C．整群抽样　　D．分层随机抽样

10．调查用的问卷中，下面的四个问题中，（　）是较好的一个问题。

A．　你和你的妈妈认为女孩几岁结婚比较好____。

B．　如果只生 1 个孩子，你希望孩子的性别是：1．女；2．男；3．随便

C．　你 1 个月工资多少_____。

D．　你一个月吃盐____克。

11．原计划调查 1000 名对象，由于种种非主观和非选择的原因，只调查到 600 名，这样的调查结果（　）。

A．　可能有偏性，因为失访者太多，可能这些失访有偏性

B．　不会有偏性，因为这种失访是自然的

C．　不会有偏性，因为这 400 名失访者不一定是某一种特征的人

D．　可能有偏性，因为 600 名对象不算多

（三）简答题

1．调查设计包含哪些内容？

2．调查表或问卷的一般结构是什么？

（四）　计算题

1．根据既往观察，人群接种某预防制剂后，体温高于 37.5℃的反应率为 10%。今欲推广使用，拟再次证实，要求容许误差在真实反应率的 20%以内，α =0.05，问按简单随机抽样需观察多少人？

2．拟用放射免疫法检测某人群血液中流行性脑脊髓膜炎特异免疫球蛋白含量，根据文献报告，其标准差约为 0.5mg/L，容许误差为 0.1mg/L，试按简单随机抽样估计样本例数。

3．表 14-3 为英格兰和威尔士男性与移民男性发病率的比较，试用直接标准化和间接标准化两种方法分别计算标准化发病率。

表 14-3　英格兰和威尔士男性与移民男性的发病率(1/10 万)

年龄分组	英格兰和威尔士			移民		
	人口(千人)	发病数	发病率	人口(千人)	发病数	发病率
0~ 4	1900	1406	74.0	26	21	80.8
5~14	3100	186	6.0	30	2	6.7
15~44	9400	1786	19.0	127	27	21.3
45~64	4900	7350	150.0	25	42	168.0
65~	2000	17400	870.0	5	48	960.0

合　计	21300	28128	132.1	213	140	65.7

4．欲检验缺铁性贫血是否是儿童智力损伤的危险因素，从“特殊”儿童日托中心选250名智力低下儿童，从正常学前教育中心选取同年龄250名儿童，测量了他们的血红蛋白等。结果见表14-4，试做OR分析。

表14-4　缺铁性贫血与儿童智力损伤关系的病例对照研究资料

缺铁性贫血	智力低下		合计
	有	无	
是	110(a)	25(b)	135(m_1)
否	140(c)	225(d)	365(m_0)
合计	250(n_1)	250(n_0)	500(n)

五、习题答案要点

（一）名词解释

1．抽样调查：抽样调查(sampling survey)是一种非全面调查，是从总体中抽取一定数量的观察单位组成样本，然后根据样本信息来推断总体特征。抽样调查是医学科研中最为常用的方法。

2．简单随机抽样：所谓简单随机抽样(simple random sampling)是在某个总体中以完全随机的方法抽取一部分个体组成样本。一般，在抽样前，需要先对抽样总体中的全部个体进行编号，然后用抽签或随机数字表的方法抽取一部分个体。

3．系统抽样：所谓系统抽样(systematic sampling)是指随机地在抽样框架内每间隔若干个个体抽取一个个体的抽样方法。在一般情况下，系统抽样的抽样误差是和简单随机抽样相仿甚至比简单随机抽样的抽样误差更小。系统抽样的抽样误差一般按简单随机抽样方法估计。

4．分层抽样：所谓分层抽样(stratified sampling)是先按对观察指标影响较大的某种特征，将总体分为若干类别(统计上称之为“层”，strata)，再从每一层内随机抽取一定数量的观察单位，合起来组成样本。分层的原则是层间差别越大越好，层内差别越小越好。

5．整群抽样：所谓整群抽样(cluster sampling)是先将总体按照某种与研究指标无关的特征化分为K个群组，每个群包括若干观察单位，然后在随机抽取k个群，将抽取的各个群的全部观察单位组成样本。

6．概率抽样：所谓概率抽样(probability sampling)就是在抽样中必须使该总体中的每一个个体都有已知的或可计算的和非零的概率被抽样抽中。常用的概率抽样方法包括：简单随机抽样、系统抽样、分层抽样和整群抽样。

7．非概率抽样：所谓非概率抽样(non-probability sampling)是指各个个体被抽样抽中的

概率是未知的和无法计算的。然而，一些非概率抽样方法，尽管不能按常规的理论来计算抽样误差和推断总体，在特定条件下，还是有用的。

8． 相对危险度：相对危险度(relative risk，RR)为暴露组发病（或死亡）率与非暴露组发病（或死亡）率之比，是队列研究中用于描述某因素与疾病发生之间的关联的主要统计学指标。

9． 病例对照研究：病例对照研究(case control study)是一种“由果推因”的回顾性观察性研究，根据有无研究疾病或其它结局，将研究人群分为病例组和对照组，追溯过去某些暴露情况，比较两组暴露水平有无差异，从而得出结局与暴露有无关联的推断。

10．队列研究：队列研究(cohort study)是一种“由因寻果”的纵向前瞻性观察研究。根据观察开始时有无暴露(exposure)史，研究者将没有出现研究疾病或其它结局(outcome)的研究人群分为暴露人群和非暴露人群，并随访观察一定时期，旨在比较两组人群的疾病‘发病”率有无差异，从而得出暴露与结局有无关联的推断。

（二） 单项选择题

1．C　2．C　3．C　4．B　5．D　6．A　7．B　8．C　9．D　10．B　11．A

（三）简答题

1．一个完整的调查设计应包括以下内容：确定明确的调查目的；确定调查对象和观察单位；确定调查方法；确定调查指标和变量；确定调查工具和设计调查表；确定样本含量；调查员的选择和培训；调查的组织计划；涉及伦理道德方面问题的处理。

2． 调查表或问卷的结构一般可划分为 4 个部分，分别为：“说明部分”、“填写说明”、“核查项目”、“调查项目”。“说明部分”主要说明调查目的，以取得调查对象的合作；“填写说明”是为了保证所有调查员和调查对象均能对调查项目及填写方法正确理解、统一认识而编写的；“核查项目”这一部分是与调查目的无关、不向调查对象询问的质量控制项目，如调查员姓名、调查日期、复核结果、未调查原因等；“调查项目”部分是调查对象填写的部分，是调查的核心内容。

（四） 计算题

1．本题为调查总体率的样本含量估计。已知：

p=0.1，d=0.2×p=0.2×0.1=0.02，双侧$u_{0.05}$=1.96

$$n=\frac{u^2_{a/2}p(1-p)}{d^2}=1.96^2\times0.1\times0.9/(0.2\times0.1)^2\approx865\text{（人）}$$

2．本题为调查总体均数的样本含量估计。已知：

s=0.5，d=0.1，双侧$u_{0.05}$=1.96

根据公式$n=(\frac{t_{a/2}s}{d})^2$计算样本含量，但在算得n之前，自由度n不能确定，$t_{0.05/2}$仍是未知的，解决的办法是先以$u_{0.05/2}$代替$t_{0.05/2}$，用迭代法求得n。

首先根据公式$n=(\frac{u_{a/2}s}{d})^2$，n=（1.96×0.5/0.1）²=96.04≈96（人）

由n=96，得自由度n=96-1=95，用n=95查t界值表得$t_{0.05/2,95}$=1.9854，再依据公式

$$n=(\frac{t_{a/2}s}{d})^2=(1.9854\times0.5/0.1)^2=98.55\approx96\text{（人）}$$

因此认为，调查样本含量为 96 人。

3． 用直接标准化计算标准化率见表 14-5。

表 14-5　直接标准化法计算移民男性的理论发病人数

年龄分组	英格兰和威尔士	移民	
	人口数	发病率（1/10万）	理论发病数
0~4	1900000	80.8	1535
5~14	3100000	6.7	208
15~44	9400000	21.3	2002
45~64	4900000	168.0	8232
65~	2000000	960.0	19200
合　计	21300000		31177

$$p' = \frac{\sum N_i p_i}{N} = \frac{31177}{21300000} \times 100000 = 146.4/10\text{万}$$

用间接标准化计算标准化率见表 14-6。

表 14-6　间接标准化法计算移民男性的理论发病人数

年龄分组	英格兰和威尔士	移民	
	发病率（1/10 万）	人口数	理论发病数
0~4	74.0	26000	19.2
5~14	6.0	30000	1.8
15~44	19.0	127000	24.1
45~64	150.0	25000	37.5
65~	870.0	5000	43.5
合　计			126.1

$$p' = P\frac{r}{\sum n_i P_i} = 132.1 \times \frac{140}{126.1} = 146.5/10\text{万}$$

4． *OR* 的计算

$$OR = \frac{110 \times 225}{25 \times 140} = 7.07$$

OR 的Mantel-Haenszel c^2 检验

H_0：缺铁性贫血与儿童智力损伤无关联，即OR 的总体参数等于 1；

H_1：OR 的总体参数不等于 1；

$c^2_{MH} = \frac{(ad-bc)^2(n-1)}{n_1 n_0 m_1 m_0} = 73.17$，$n = 1$

73.17> $c^2_{0.05,1}$=7.88，P<0.05，接受H_1。

故可认为缺铁性贫血与儿童智力损伤有关联。智力低下儿童患有缺铁性贫血的危险是正常同龄儿童的7倍。

（颜虹　姜建辉）

第十五章　医学人口统计与疾病统计常用指标

一、教学大纲要求

（一） 掌握内容

1．医学人口统计常用统计指标的意义及用途

（1）人口数与人口构成常用指标：人口总数、性别比、老年人口系数、少年儿童人口系数；

（2）人口金字塔；

（3）生育与计划生育常用指标：粗出生率、总生育率、终生生育率、总和生育率、自然增长率；

（4）死亡统计常用指标：粗死亡率、年龄别死亡率、新生儿死亡率、婴儿死亡率、5 岁以下儿童死亡率、标准化死亡率、死因别死亡率、死因顺位。

2．疾病统计常用统计指标的意义及用途

发病率、患病率、病死率、治愈率、生存率。

（二） 熟悉内容

医学人口统计和疾病统计的其它指标。

（三） 了解内容

医学人口统计和疾病统计的含义及其资料来源。

二、教学内容精要

(一) 医学人口统计常用指标的意义及其用途

1． 人口数与人口构成常用指标

（1）人口数：人口数(population) 又称人口总数，一般指一个国家或地区某一特定时间点的人口数。通过一次人口普查，可得较好的人口数统计。根据我国的户籍登记，也可获得户籍人口数。在人口流动较多的情况下，还可按居住地来统计人口数。

（2）性别比：以女性人口为 100，计算男女性人口数之比，称为性别比或性比例。

$$\text{性别比}=\frac{\text{男性人数}}{\text{女性人数}}\times 100 \qquad (15\text{-}1)$$

（3）老年人口系数：指老年人口在总人口中所占的比重，是说明人口老龄程度的指标，可作为划分人口类型的尺度。

$$\text{老年人口系数}=\frac{\text{65 岁及以上人口数}}{\text{人口总数}}\times 100\% \qquad (15\text{-}2)$$

（4）少年儿童人口系数：指少年儿童人口在总人口中所占的比重，是划分人口类型的指标之一。

$$\text{少年儿童人口系数}=\frac{\text{14岁及以下人口数}}{\text{人口总数}}\times 100\% \qquad (15\text{-}3)$$

2．人口金字塔

（1）人口金字塔：人口金字塔(pyramid)**是一种用几何图形来形象的表示人口性别年龄构成的方法。将人口的性别、年龄分组数据，以年龄（或出生年份）为纵轴，以人口数或年龄构成比为横轴，按左侧为男、右侧为女绘制的直方图，其型如金字塔，称为人口金字塔。人口金字塔更形象直观地反映了人口的年龄性别构成，便于说明和分析人口的现状、类型。**

（2）人口金字塔的类型：人口金字塔可分为三种类型：年轻型、成年型和年老型。它们的形状各不相同。年轻型：塔顶尖、塔底宽。成年型：塔顶、塔底宽度基本一致，在塔尖处才逐渐收缩。年老型：塔顶宽，塔底窄。

从人口年龄结构对今后人口增长速度影响的角度，又可将人口金字塔分为增长型、静止型和缩减型，分别与年轻型、成年型和年老型相对应。

3．生育与计划生育常用指标

（1）粗出生率：粗出生率（crude birth rate，*CBR*）又称出生率，指某地某年平均每千人口中的出生数（活产数），人口的出生率明显受人口的性别年龄结构的影响。其算式为：

$$\text{粗出生率}=\frac{\text{某年活产总数}}{\text{同年平均人口数}}\times 1000‰ \qquad (15\text{-}4)$$

（2）总生育率：总生育率(general fertility rate，*GFR*)**又称生育率，指某地某年平均每千名育龄妇女的活产数，是测量人群生育水平的指标。其算式为：**

$$\text{生育率}=\frac{\text{某年活产总数}}{\text{同年15~49岁妇女平均人口数}}\times 1000‰ \qquad (15\text{-}5)$$

（3）终生生育率：终生生育率（life-time fertility rate，*LTFR*）说明一批经历过整个育龄期的妇女一生的生育水平。终生生育率由于观察时间很长，一般很难观察到。

$$\text{终生生育率}=\frac{\text{某批妇女生育的活产子女数}}{\text{经历过整个育龄期的该批妇女数}}\times 1000‰ \qquad (15\text{-}6)$$

（4）总和生育率：总和生育率（total fertility rate，*TFR*）假定一批妇女按某一套年龄别生育率计算，平均在整个育龄期会有几个活产。

该指标反映的是调查年时间横断面上的生育水平。因其消除了年龄构成不同对生育水平的影响，故不同地区、不同年度的总和生育率可以直接比较，因而应用较广，是较好的测量生育水平的指标。

$$\text{总和生育率}=\sum(\text{年龄组组距}\times\text{各年龄组生育率}) \qquad (15\text{-}7)$$

（5）自然增长率：自然增长率（natural increase rate, *NIR*）**为粗出生率与粗死亡率之差，是测量人口再生育的指标。易受人口性别、年龄的影响，只能粗略的估计人口的一般增长趋势，不能用来估计未来人口的发展速度。**

$$\text{人口自然增长率}=\text{粗出生率}-\text{粗死亡率} \qquad (15\text{-}8)$$

4．死亡统计常用指标

（1）粗死亡率：粗死亡率(crude death rate，*CDR*)又称死亡率（death rate），是某时期（一般是1年）死亡总数除以该时期的平均人口数或期中人口数所得的商。如果用一年的资料计算年死亡率，分子是一年内的死亡数，分母就是该年的平均人口数或年中人口数。粗死亡率说明人群中总的死亡水平，易受人口性别、年龄的影响。

$$\text{粗死亡率}=\frac{\text{某年死亡数}}{\text{同年平均人口数}}\times 1000‰ \qquad (15\text{-}9)$$

（2）年龄别死亡率：年龄别死亡率（age-specific death rate，*ASDR*）指某年某年龄别平均每千人口中的死亡数。

$$\text{年龄别死亡率}=\frac{\text{某年某年龄组死亡人数}}{\text{同年该年龄组平均人口数}}\times 1000‰ \qquad (15\text{-}10)$$

（3）标准化死亡率：一群人的死亡率高低受该人群年龄构成的影响，所以不同人群或同一人群不同时间的死亡率比较时，应该考虑用某种方法消除年龄构成的影响。标准化死亡率（standardized mortality rate，*SMR*）就是这样的一个指标。直接法计算的标准化死亡率，就是用同一套标准的年龄构成比与各自的年龄组死亡率乘积的总和。

（4）婴儿死亡率：婴儿死亡率（infant mortality rate，*IMR*）指某地某年不满一周岁婴儿的死亡数与同期活产总数的比值。婴儿死亡率的高低对平均寿命有重要的影响，它是反映社会卫生状况和婴儿保健工作的重要指标，也是死亡统计指标中较为敏感的指标。

$$\text{婴儿死亡率}=\frac{\text{某年不满周岁婴儿死亡数}}{\text{同期活产数}}\times 1000‰ \qquad (15\text{-}11)$$

（5）新生儿死亡率：新生儿死亡率（neonatal mortality rate，*NMR*）指某地某年未满28天的新生儿的死亡数与同期活产总数的比值。与婴儿死亡率同样是反映妇幼卫生工作的重要指标。新生儿死亡数在婴儿死亡数中占很大的比重（约占 50%），因此，降低新生儿死亡率是降低婴儿死亡率的关键。但是，新生儿死亡漏报现象非常严重。在我国，有的边远地区新生儿死亡漏报率高达100%。新生儿死亡漏报直接影响到该指标的准确性。

$$\text{新生儿死亡率}=\frac{\text{某年不满28天新生儿死亡数}}{\text{同期活产数}}\times 1000‰ \qquad (15\text{-}12)$$

（6）5岁以下儿童死亡率

由于儿童死亡率比较高，且不易获得完整的统计资料，在卫生事业不发达或统计制度不健全的国家和地区，婴儿和新生儿死亡数往往有漏报。故也常用5岁以下儿童死亡率来反映婴幼儿的死亡水平。

$$5\text{岁以下儿童死亡率}=\frac{\text{某年不满5岁儿童死亡数}}{\text{同年活产数}}\times 1000‰ \qquad (15\text{-}13)$$

（7）死因别死亡率：死因别死亡率（cause-specific death rate，*CSDR*）指因某种原因（疾病）所致的死亡率，是死因分析的重要指标，反映各类病伤死亡对居民生命的危害程

度。

$$某死因死亡率=\frac{某年内某种原因的死亡人数}{同年平均人口数}\times 100000/10万 \quad (15\text{-}14)$$

（8）死因顺位：指按各类死因构成比的大小或死因别死亡率的高低顺序，由高到低排列的位次，说明各类死因的相对重要性。死因顺位可以反映各种死因所致死亡的相对重要性。

(二) 疾病统计常用统计指标

1．发病率：发病率（incidence rate）表示在观察期内，可能发生某种疾病的一定人群中新发该病的频率。

$$某病发病率=\frac{观察期内新发生某病的例数}{同期平均人口数}\times 1000‰ \quad (15\text{-}15)$$

2．患病率：一般所说的患病率(prevalence rate)，又称现患率，指时点患病率(point prevalence rate)，是某一时间横断面上某病患者数占受检人数的比例，它是一种静态指标，虽然名称是率，但它的性质是比例。通常用于描述病程较长的慢性病或发病时间不易明确的疾病的患病情况。

$$患病率=\frac{现患人数}{受检人数}\times 1000‰ \quad (15\text{-}16)$$

在某些场合，也使用时期患病率（period prevalence rate），时期患病率的分子实际上是该时期起始点的患病例数与整个时期的新病例数之和，分母是同期平均人口数。

3．某病病死率：某病病死率 (fatality rate)表示在规定的观察期内，某病患者中因该病而死亡的频率。

$$某病病死率=\frac{观察期内因某病死亡的人数}{同期该病患者数}\times 1000‰ \quad (15\text{-}17)$$

4．某病死亡率：某病死亡率（mortality rate）表示在规定的观察期内，人群中因某病而死亡的频率。它可以反映不同地区或年代某种疾病的死亡水平。

$$某病死亡率=\frac{观察期内因某病死亡的人数}{同期平均人口数}\times 1000‰ \quad (15\text{-}18)$$

5．治愈率：治愈率（cure rate）指受治病人中治愈的频率。主要适用于一些急性病的疗效统计。

$$治愈率=\frac{治愈人数}{受治人数}\times 100\% \quad (15\text{-}19)$$

6．生存率：生存率（survival rate）是指观察对象能存活到某一时点的概率。常用的是一年生存率、五年生存率和十年生存率等。临床上，一些慢性病的病人经过某种治疗后的治疗效果，常用n年生存率来表示。对恶性肿瘤等疾病，难说“治愈”，用n年生存率来表示治疗效果或凶险程度是比较合适的。

$$n年生存率=\frac{活满n年的例数}{观察例数}\times 100\% \quad (15\text{-}20)$$

生存率一般要用寿命表法（即Kaplan-Meier 法）计算。不宜按照对上述公式的直观理解，用“直接法”进行计算。

（三）医学人口统计的含义及其资料来源

1．医学人口统计：是应用人口统计学的理论与方法，从人类健康和卫生保健的角度研究人口的数量、结构、变动及其与卫生事业发展的相互关系，是人口统计学在居民健康和卫生保健领域中的应用，是卫生统计学的重要组成部分。

2．资料来源：主要来源于人口统计收集的资料，有以下几个方面：

（1） 人口普查；

（2） 人口抽样调查；

（3） 人口登记，包括生命事件登记（出生、死亡、胎儿死亡、结婚、离婚、收养、生育、认领、离弃等）、人口迁移变动登记和户口登记。

（四）疾病统计的意义及其资料来源

4． 疾病统计：是居民健康统计的重要内容之一，它的任务是研究疾病在人群中发生、发展及其流行的规律，为病因学研究、疾病防治和评价疾病防治效果提供科学依据。

5．资料来源：主要来源于以下三个方面：

（1）疾病报告和报表资料；

（2）医疗卫生工作记录；

（3）疾病调查资料。

三、典型试题分析

（一）名词解释

婴儿死亡率。

答案：婴儿死亡率（infant mortality rate, *IMR*）指某地某年不满一周岁婴儿的死亡数与同年活产总数的比值。婴儿死亡率的高低对平均寿命有重要的影响，它是反映社会卫生状况和婴儿保健工作的重要指标，也是死亡统计指标中较为敏感的指标。其计算式为：

$$\text{婴儿死亡率}=\frac{\text{某年不满周岁婴儿死亡数}}{\text{同年活产总数}}\times 1000‰$$

婴儿死亡率的高低对平均寿命有重要的影响，它是反映社会卫生状况和婴儿保健工作的重要指标，也是死亡统计指标中较为敏感的指标。

[评析] 本题考点：婴儿死亡率概念的理解。

（二）单项选择题

1．在死因统计分析中，死因顺位是按（ ）的高低顺序，由高到低排列的位次。

A．发病率　　B．死因百分构成比或死因别死亡率

C．死因别病死率　　D．患病率

答案：B。

[评析] 本题考点：对死因顺位含义的理解。

死因顺位是指按各类死因构成比的大小或死因别死亡率的高低顺序，由高到低排列的位次。死因顺位可以反映各种死因所致死亡的相对重要性。

2．反映疾病发生频度的指标有（　）。

A．患病率、感染率　　B．发病率、病死率

C．感染率、发病率　　D．发病率、患病率

答案：D。

[评析] 本题考点：反映疾病发生频度的指标。

发病率与患病率同为疾病发生频度的指标。发病率表示一定时期内，在特定人群中新发生的某病病例数，反映某病新发病例的发生频度。患病率是指某时点上受检人数中现患某种疾病的人数，通常用于描述病程较长的慢性病或发病时间不易明确的疾病的患病情况。

3．总和生育率是指（　）。

A．一批妇女一生平均生育的子女数

B．一批妇女按某年的年龄别生育水平计算，一生平均生育的子女数

C．一批妇女某年的平均活产数

D．某年龄段的育龄妇女某年的平均活产数

答案：B。

[评析] 本题考点：总和生育率概念的理解。

总和生育率（total fertility rate，*TFR*）：假定一批妇女按某一套年龄别生育率计算，平均在整个育龄期会有几个活产。计算公式为：

$$\text{总和生育率}=\sum(\text{年龄组组距}\times\text{各年龄组生育率})$$

该指标反映的是调查年时间横断面上的生育水平。因其消除了年龄构成不同对生育水平的影响，故不同地区、不同年度的总和生育率可以直接比较，因而应用较广，是较好的测量生育水平的指标。

4．人口金字塔可以用来反映（　）。

A．人口出生情况　　B．人口死亡情况

C．人口的年龄性别构成情况　　D．人口迁入迁出情况

答案：C。

[评析] 本题考点：人口金字塔的意义及用途。

人口金字塔是将人口的性别、年龄分组数据，以年龄（或出生年份）为纵轴，以人口数或年龄构成比为横轴，按左侧为男、右侧为女绘制的直方图，其型如金字塔，故称为人口金字塔。人口金字塔更形象直观的反映了人口的年龄性别构成，便于说明和分析人口的现状和类型。

5．老年人口比重增大，可使（　）。

A．粗死亡率增高　　B．粗死亡率下降

C．婴儿死亡率下降　　D．出生率迅速下降

答案：A。

[评析]本题考点：粗死亡率的概念及其影响因素。

粗死亡率又称死亡率，是某时期（一般是 1 年）死亡总数除以该时期的平均人口数或期中人口数所得的商。如果用一年的资料计算年死亡率，分子是一年内的死亡数，分母就是该年的平均人口数或年中人口数。粗死亡率说明人群中总的死亡水平，易受人口性别、年龄的影响。一般情况下，老人和婴儿的死亡率较高，男性死亡率高于女性。计算公式为：

$$\text{粗死亡率}=\frac{\text{某年死亡人数}}{\text{同年平均人口数}}\times 1000‰$$

（三）简答题

何谓人口老龄化？请简述其影响因素。

答案：人口老龄化是指老年人口在人口中所占的比重升高的现象。在没有迁移的情况下，人口老龄化的进程主要受生育率和死亡率两种因素的影响。死亡率（主要是中老年人口的死亡率）降低，使寿命延长，老年人口比重增加。生育率下降，使低年龄人口的比重降低，高年龄人口的比重相应增加。一般来说，人口老龄化的速度和程度主要取决于生育率的下降速度。当生育率水平下降达到很低的水平且很难再有较大程度的降低时，中老年人口死亡率的降低对人口老龄化的影响才比较明显。

（四）计算题

表 15-1 是某地区的人口学调查资料，请就此资料作如下分析：

1．计算全人口的性别比；

2．计算育龄期妇女（15-49 岁）占总人口的百分比；

6．计算负担系数；

7．计算老龄人口的比重。

表 15-1　某地男、女性人口占总人口的百分比

年龄组（岁）	男	女	年龄组（岁）	男	女
0～	4.2	4.0	45～	2.4	2.7
5～	3.2	3.1	50～	2.1	2.4
10～	4.4	4.2	55～	1.2	2.2
15～	5.5	5.3	60～	1.3	2.4
20～	5.1	5.2	65～	1.1	1.4
25～	6.0	6.1	70～	0.8	1.2
30～	4.3	4.5	75～	0.5	0.9
35～	3.2	3.3	80～	0.2	0.5
40～	2.3	2.5	85～	0.1	0.2

解：1．计算全人口的性别比

$$\text{全人口的性别比}=\frac{\text{男性人口数}}{\text{女性人口数}}\times 100$$

$$=\frac{\text{男性人口占全人口的百分比}}{\text{女性人口占全人口的百分比}}\times 100$$

$$=49.9/52.1\times 100=91.94$$

2．计算育龄妇女占总人口的百分比

$$\text{育龄妇女占总人口的百分比}=\frac{\text{育龄期妇女人数}}{\text{总人口数}}\times 100\%$$

$$=(5.3+5.2+6.1+4.5+3.3+2.5+2.7)\times 100\%$$

$$=29.6\%$$

3．计算负担系数

负担系数又称抚养比和抚养系数，是指人口中非劳动年龄人数与劳动年龄人数之比。

$$\text{总负担系数}=\frac{\text{14岁及以下人口数+65岁及以上人口数}}{\text{15～64岁人口数}}\times 100\%$$

$$=30.0/70.0\times 100\%=42.86\%$$

4．计算老年人口系数

$$\text{老年人口系数}=\frac{\text{65岁及以上人口数}}{\text{人口总数}}\times 100\%$$

$$\text{老年人口系数}=\frac{\text{65岁及以上各年龄组人口百分比之和}}{100}=6.9\%$$

[评析]本题考点：人口调查资料的统计分析。

人口普查或抽样调查获得的人口资料分析，往往是从人口的基本特征、人口年龄构成、性别比及人口金字塔等诸方面进行描述，计算其相应的统计指标，以反映人口的数量、结构及变动情况。

四、习　题

（七）名词解释

1.老年人口系数　2.负担系数　3.人口金字塔　4.出生率　5.总和生育率

6.标准化死亡率　7.计划生育率　8.死因别死亡率　9.孕产妇死亡率　10.生存率

（八）单项选择题

1．出生率下降，可使（　）。

A．婴儿死亡率下降　　B．老年人口比重增加

C．总死亡数增加　　D．老年人口数下降

2．计算某年婴儿死亡率的分母为（　）。

A．年活产总数　　B．年初0岁组人口数

C．年中0岁组人口数　　D．年末0岁组人口数

3．自然增长率是估计一般人口增长趋势的指标，它的计算是（　）。

A．出生数 — 死亡数　　B．粗出生率 — 粗死亡率

C．标化出生率 — 标化死亡率　　D．年末人数 — 年初人数

4．计算某年围产儿死亡率的分母是（　）。

A．同年妊娠28周以上的妇女数

B．同年妊娠28周以上出生的活产数

C．同年死胎数 + 死产数 + 活产数

D．同年出生后7天内的新生儿数

5．终生生育率是指（　）。

A．一批经历过整个育龄期的妇女一生平均生育的子女数

B．一批妇女按某时的生育水平，一生可能生育子女数

C．一批经历过整个育龄期的妇女某年的平均活产数

D．某年龄段的妇女某年的平均活产数

6．年龄别生育率是指（　）。

A．每1000名妇女一生平均生育的子女数

B．每1000名妇女按某时的生育水平，一生可能生育子女数

C．每1000名妇女某年的平均活产数

D．每1000名某年龄段的育龄妇女某年的活产数

7．婴儿死亡率是指（　）。

A．0岁死亡率

B．活产婴儿在生活一年内的死亡概率

C．某年不满1岁婴儿死亡数与同年活产总数之比

D．某年不满1岁婴儿死亡数与同年婴儿总数之比

8．某病病死率和某病死亡率均为反映疾病严重程度的指标，两者的关系为（　）。

A．病死率高，死亡率一定高

B．病死率高，死亡率不一定高

C．青年人口中，病死率高，死亡率也高

D．女性人口中，病死率高，死亡率也高

9．总和生育率下降，可使老年人口百分比（　）。

A．上升　　B．下降

C．毫无关系　　D．以上答案均不对

10．观察某种疫苗的预防效果，若第一季度初接种了400人，第二季度初接种了300人，第三季度初接种了100人，第四季度初接种了200人，到年终总结，这1000人中发病者20人，计算发病率的分母应该是（　）。

A．1000人　　B．（400+200）/2 人

C．（400+300+100+200）/4人　　D．400+300×3/4 +100×1/2 + 200×1/4 人

11．随访观察某种慢性病1000人的治疗结果，第一年死了100人，第二年死了180人，第三年死了144人，则该慢性病的3年生存率的算法为（　）。

A．（0.9 + 0.8 + 0.8）/3　　B．1－0.10×0.20×0.20）

C．1－0.10－0.20－0.20　　D．0.90×0.80×0.80

12．老年人口一般是指（　）。

A．50岁及以上的人口　　B．55岁及以上的人口

C．60岁及以上的人口　　D．65岁及以上的人口

（三）简答题

1．发病率、时点患病率、时期患病率的区别。

2．疾病统计的观察单位“病人”和“病例”的区别。

3．病死率和死亡率的区别。

五、习题答题要点

（一）名词解释

1． 老年人口系数：老年人口系数指老年人口在总人口中所占的比重，是说明人口老年化程度的指标，可作为划分人口类型的尺度。一般把 65 岁及以上的人口称为老年人口，而发展中国家倾向于以 60 岁作为老年年龄界限。老年人口系数的算式为：

$$\text{老年人口系数} = \frac{\text{65 岁及以上的人口数}}{\text{人口总数}} \times 100\%$$

2.负担系数：负担系数又称抚养比或抚养系数，是指人口中非劳动年龄人数与劳动年龄人数之比。一般以 14～64 岁为劳动年龄，14 岁及以下和 65 岁及以上为非劳动年龄或抚养年龄。负担系数包括三个指标：总负担系数、少年儿童负担系数和老年负担系数。各国由于人口年龄构成不同，负担系数也有所不同。

3.人口金字塔：将人口的性别、年龄分组数据，以年龄（或出生年份）为纵轴，以人口数或年龄构成比为横轴，按左侧为男、右侧为女绘制的直方图，其型如金字塔，称为人口金字塔(pyramid)。人口金字塔更形象直观的反映了人口的性别年龄构成，便于说明和分析人口的现状和类型。

4.出生率：出生率（birth rate，*BR*）又称粗出生率，指某地某年平均每千人口中的出生数（活产数），人口的出生率明显受人口的性别、年龄结构和婚姻状况的影响，因此，它只能粗略的反应生育水平。其算式为：

$$\text{出生率} = \frac{\text{某年活产总数}}{\text{同年平均人口数}} \times 1000‰$$

5.总和生育率：总和生育率（total fertility rate，*TFR*）假定一批妇女按某一套年龄别生育率计算，平均在整个育龄期会有几个活产。

该指标反映的是调查年时间横断面上的生育水平。因其消除了年龄构成不同对生育水平的影响，故不同地区、不同年度的总和生育率可以直接比较，因而应用较广，是较好的测量生育水平的指标。

$$\text{总和生育率} = \sum \text{（年龄组组距×各年龄组生育率）}$$

6．标准化死亡率：一群人的死亡率高低受该人群年龄构成的影响，所以不同人群或同一人群不同时间的死亡率比较时，应该考虑用某种方法消除年龄构成的影响。标准化死亡率（standardized mortality rate，*SMR*）就是这样的一个指标。直接法计算的标准化死亡率，就是用同一套标准的年龄构成比与各自的年龄组死亡率乘积的总和。

7.计划生育率：计划生育率是指每 1000 名活产中符合计划生育要求者的例数。他综合说明计划生育的质量，可与反映计划生育工作的其他指标联合，用于评价计划生育工作。

$$\text{计划生育率} = \frac{\text{某年符合计划生育的活产数}}{\text{同年活产总数}} \times 100\%$$

8.死因别死亡率：死因别死亡率（cause-specific death rate）指因某种原因（疾病）所致的死亡率。其算式为：

$$某死因死亡率=\frac{某年某死因死亡人数}{同年平均人口数}\times 100000/10万$$

死因别死亡率是死因分析的重要指标，它反映各类病伤死亡对居民生命健康的危害程度。

9.孕产妇死亡率：孕产妇死亡率（maternal mortality rate）指某年中由于怀孕和分娩及其并发症造成的孕产妇死亡人数与同年活产数之比，以万分率或十万分率表示，其算式为：

$$孕产妇死亡率=\frac{某年孕产妇死亡人数}{同年活产总数}\times 100000/10万$$

孕产妇死亡率不仅可以评价妇女保健工作，而且间接反映一个国家的卫生文化水平。

10.生存率：生存率（survival rate）是指观察对象能存活到某一时点的概率。常用的是一年生存率、五年生存率和十年生存率等。临床上，一些慢性病的病人经过某种治疗后的治疗效果，常用n年生存率来表示。对恶性肿瘤等疾病，难说“治愈”，用n年生存率来表示治疗效果或凶险程度是比较合适的。

$$n年生存率=\frac{活满n年的例数}{观察例数}\times 100\%$$

生存率一般要用寿命表法（即Kaplan-Meier法）计算。不宜按照对上述公式的直观理解，用“直接法”进行计算。

（二）单项选择题

1.B　2.A　3.B　4.C　5. A　6. D　7. C　8. B　9. A　10. D
11. D　12. D

（三）简答题

1．发病率、时点患病率、时期患病率的区别。

（1）发病率是指观察期内，可能发生某病的人群中新发病例的频率，其观察期多为年、月、日等，急性常见病多计算发病率。

（2）时点患病率反映在检查或调查时点一定人群中某病的现患情况（包括该病的新旧病例数）。观察时点在理论上是无长度的，但实际上观察时间不宜过长，一般不超过个月。

（3）时期患病率反映在观察期间一定人群中存在或流行某病的频度，包括观察期间的新发病例和现患病例数，常为慢性病的统计指标，但收集资料很困难。

2．疾病统计的观察单位“病人”和“病例”的区别。

（1）一个人每次患病都可作为一个病例。以病例为单位的疾病统计，可研究居民各种疾病的频度、疾病的种类及疾病的变动，以获得居民患病的基本规律。

（2）病人是指一个有病的人。在观察期间内，观察对象患有疾病即算作一个病人，不管其患病的种类及患病次数的多少。以病人为单位的疾病统计，在一定程度上反映居民的患病频度，可找出具体的患病人群，便于开展对病人个人的防治工作。

3．病死率和死亡率的区别。

（1）　某病病死率表示在规定的观察期内，某病患者中因该病而死亡的频率。它是反

映疾病的严重程度的指标。在用病死率进行比较时应注意内部构成不同的影响。计算公式为：

$$某病病死率=\frac{观察期内因某病死亡的人数}{同期该病患者数}\times 1000‰$$

（2）某病死亡率表示在规定的观察期内，人群中因某病而死亡的频率。它可以反映不同地区或年代某种疾病的死亡水平。计算公式为：

$$某病死亡率=\frac{观察期内因某病死亡的人数}{同期平均人口数}\times 1000‰$$

（詹绍康　王霞）

第十六章　寿命表

一、教学大纲要求

（一）掌握内容

1．寿命表的概念。

2．寿命表的分类：现时寿命表、定群寿命表；完全寿命表、简略寿命表等。

3．寿命中的各项指标：年龄、年龄组死亡概率、尚存人数与死亡人数、生存人年数、平均预期寿命。

4．寿命表的编制：简略寿命表的编制、去死因寿命表的编制。

5．寿命表的分析：寿命表的指标分析；寿命表的应用。

（二）熟悉内容

全死因寿命表、定群寿命表的编制方法。

（三）了解内容

寿命表在生存及死亡分析中的应用。

二、教学内容精要

(一) 寿命表的概念

寿命表（life table）是根据特定人群的年龄组死亡率编制出来的一种统计表。寿命表的指标可以用来评价居民的健康状况。寿命表的编制需要完整的人口资料与死亡资料。

寿命表的分类：现时寿命表（current life table**）和定群寿命表（**cohort life table**）。**

现时寿命表是指从一个断面看问题，假定有同时出生的一代人，按照某种人群现时人口实际年龄组死亡率陆续死去，计算出这一代人按年龄的一系列指标。依据年龄分组不同，现时寿命表可分为完全寿命表（年龄分组的组距是 1 **岁）和简略寿命表（年龄分组的组距一般是** 5 **岁）。其中简略寿命表更常用。**

定群寿命表是指对某特定的人群中的每一个人，从进入该特定人群直到最后一个人死亡，记录的实际死亡过程。因为人的生命周期很长，这种方法实现起来难度很大，因此一般来说应用于涉及事物寿命现象的问题，不一定是人群从出生到死亡的过程。

(二) 寿命表的编制原理与方法

1．**年龄　寿命表中的年龄是指“刚满年龄”**(exact age)

2．**年龄组死亡概率（**age specific probability of dying**）　是指** X **岁尚存者在今后一年或** n **年内死亡的可能性。它和年龄组死亡率不是一个概念。在编制寿命表时，这是一个很关键的指标。**

$$q_X = \frac{d_X}{l_X} \quad \text{或} \quad {}_nq_X = \frac{{}_nd_X}{l_X} \tag{16-1}$$

其中 q_X **表示** X **岁尚存者在今后一年内的死亡概率；**${}_nq_X$ **表示** X **岁尚存者在今后** n **年的死**

亡概率；d_X 表示寿命表死亡人数；${}_n d_X$ 表示在X～（X+n）岁期间的寿命表死亡人数。

3．尚存人数与死亡人数（number of survival person-years） 尚存人数l_X表示同时出生的一代人中活满X岁的人数。

尚存人数l_X，死亡人数d_X（${}_n d_X$）及死亡概率q_X（${}_n q_X$）关系如下：

$$d_X = l_X \cdot q_X \quad \text{或} \quad {}_n d_X = l_X \cdot {}_n q_X \tag{16-2}$$

$$l_{X+1} = l_X - d_X \quad \text{或} \quad l_{X+n} = l_X - {}_n d_X \tag{16-3}$$

4．生存人年数（number of survival person-years）及生存总人年数（total number of survival person-years） X岁尚存者在今后一年（n年）内的生存人年数L_X（${}_n L_X$），即l_X曲线下，X～（$X+n$）间的面积。这个面积近似梯形面积。但婴儿组的人年数及最后一组的人年数用下面公式计算：

婴儿组 $$L_0 = l_1 + a_0 \times d_0 \tag{16-4}$$

其中a_0是指0岁组死亡者的平均存活年数。

最后一个年龄组 $$L_w = \frac{l_w}{m_w} \tag{16-5}$$

其中L_w是最后一个年龄组的生存人年数；l_w是指尚存人数；m_w是指死亡统计中的最后一组死亡率。

5．平均预期寿命（life expectancy） 表示X岁尚存者预期平均尚能存活的人年数。

$$e_x = \frac{T_x}{l_x} \tag{16-6}$$

（三）简略寿命表

简略寿命表(abridged life table)一般以日历年度的人口资料为依据，统计数字的准确与否，直接影响寿命表指标的准确性与可靠性，因此必须要求准确的数据资料。简略寿命表习惯上组距是5岁，但零岁作为一个独立的组。由于简略寿命表年龄分组少，每个年龄组人口数较多，年龄组死亡率较稳定，卫生统计中比较常用。

（四）去死因寿命表

去死因寿命表（cause eliminated life table）是用来分析某种疾病或某类疾病对平均预期寿命等指标的影响，可以综合说明某类死因对人群生命的影响程度，它不受人口年龄结构的影响，而且它既能说明某类死因对全人口的综合作用，又能表达对某年龄组人口的作用。

去死因寿命表的编制方法 去死因寿命表中各项指标的意义与全死因寿命表相同。编制去某死因寿命表的关键是求去某死因后各年龄组生存率（${}_n p_X^{-i}$），有了${}_n p_X^{-i}$，就可以仿照编制全死因寿命表的方法，编制去某死因寿命表，其中

$${}_n p_X^{-i} = ({}_n p_X)^{n^{r_x^{-i}}} \tag{16-7}$$

（五）寿命表的分析与应用

1．寿命表的分析 寿命的各项指标l_X、${}_n d_X$、${}_n q_X$、e_X都用来评价居民的健康水平。其中最主要的指标是平均预期寿命。

寿命表尚存人数：反映在一定年龄组死亡率基础上，一代人口的生存过程，一般用线图表示。尚存人数随年龄增加而减少。

寿命表死亡人数：反映在一定年龄组死亡基础上，一代人口的死亡过程。一般用直方图表示。横坐标为年龄，纵坐标为死亡人数。

寿命表死亡概率：取决于各年龄组死亡率，一般用半对数线图表示。

预期寿命：预期寿命是评价居民健康状况的主要指标。一般用线图表示。

2．寿命表的应用

寿命表主要应用于：(1）评价国家或地区居民健康水平。(2）利用寿命表研究人口再生产情况。(3）利用寿命表指标进行人口预测。(4）利用寿命表方法研究人群的生育、发育及疾病的发展规律。

三、典型试题分析

（一）名词解释

平均预期寿命。

平均预期寿命（life expectancy）：寿命表平均预期寿命是指 X 岁尚存者预期平均尚能存活的年数。平均预期寿命是评价居民健康状况的主要指标。刚满 X 岁者的平均预期寿命受 X 岁以后各年龄组死亡率的综合影响。

（二）单项选择题

某地某年女性简略寿命表中0岁组的预期寿命是65.5岁,则1岁组的预期寿命为(　　)。

G． 等于65.5岁

H． 小于65.5岁

I． 大于65.5岁

D． 不一定

答案：D

[评析] 本题考点： 0岁组的预期寿命与1岁组预期寿命的关系。

0 岁组的预期寿命简称平均寿命，它是各年龄组死亡率的综合反映，任何一个年龄组的死亡水平发生变化，都会引起平均寿命的改变，但婴儿死亡率对平均寿命的影响更为明显。一般来说，随着年龄的增长，预期寿命应逐渐下降，0 岁组的预期寿命应高于 1 岁组预期寿命，但是当婴儿死亡率较高时，就会出现 0 岁组的预期寿命应低于 1 岁组预期寿命的现象。

四、习　题

（十六）　名词解释

1．寿命表　2．现时寿命表　3．完全寿命表　4．简略寿命表 5．定群寿命表

6．年龄组死亡概率　7．尚存人数

（十七）　单项选择题

1． 在寿命表中，若 X 岁到 X+1 岁的死亡概率为 ${}_1q_x$，X+1 到 X+2 的死亡概率 ${}_1q_{x+1}$，则 X 到 X+2 的死亡概率为 (　)。

A．${}_1q_x \times {}_1q_{x+1}$　　B．$1-{}_1q_x \times 1q_{x+1}$

C．$(1-{}_1q_x) \times (1-{}_1q_{x+1})$　　D．$1-(1-{}_1q_x) \times (1-{}_1q_{x+1})$

2．卫生统计学中目前常用的计算某年婴儿死亡率的分母是 (　)。

A．年初 0 岁组人口数　　B．年中 0 岁组人口数

C．年末 0 岁组人口数　　D．年出生数

（十八） 简答题

1．年龄组死亡率与寿命表死亡概率有什么区别和联系？

2．平均寿命与平均死亡年龄的区别？

3．简略说明寿命表中的$m_{85(+)}$与e_{85}的关系。

（四）计算题

1．下表为某市 1998 年男性居民的按年龄分组的生存资料，试编制简略寿命表。

表 16-1 某市 1998 年男性居民的按年龄分组的生存资料

年龄组（岁）	平均人口数	实际死亡人数	年龄组（岁）	平均人口数	实际死亡人数
0~	18753	246	40~	56806	134
1~	54325	60	45~	65863	239
5~	64063	46	50~	54243	346
10~	94683	64	55~	43355	528
15~	114332	90	60~	32004	763
20~	126941	123	65~	24445	972
25~	118930	127	70~	12818	897
30~	91922	104	75~	5813	647
35~	62290	92	80~	2685	517

注：a_0 =0.145

五、习题答题要点

（十七） 名词解释

1．寿命表：寿命表（life table**）亦称生命表，是根据特定人群的年龄组死亡率编制出来的一种统计表。寿命表中各项指标不受人口年龄构成的影响，不同人群的寿命表指标具有良好的可比性。**

2．现时寿命表：现时寿命表（current life table**）指从一个断面看问题，假定有同时出生的一代人，按照某种人群现时人口实际年龄组死亡率陆续死去，计算出这一代人按年龄的一系列指标。**

3．完全寿命表：在编制寿命表时，如果年龄分组的组距是一岁，则称为完全寿命表（complete life table**），编制完全寿命表时观察人数要足够多。**

4．简略寿命表：如果年龄分组的组距不是一岁时，刚称为简略寿命表（abridged life table**），简略寿命表的组距一般是 5 岁，但零岁作为一个独立组。**

5．定群寿命表：定群寿命表（cohort life table**）亦称队列寿命表，它是对某特定的人群中的每一个人，从进入该特定人群直到最后一个人死亡，记录的实际死亡过程。**

6．年龄组死亡概率：年龄组死亡概率（age specific probability of dying**）是指 X 岁尚存者在今后一年或 n 年内死亡的可能性。它和年龄组死亡率不是一个概念。**

7．尚存人数：寿命表尚存人数（number of survivors）是指同时出生的一代人中活满 X 岁的人数。

（二）单项选择题

1．D　　2．D

（三）简答题

1．二者的区别：年龄组死亡率是说明某年龄组人口在一年内实际的死亡水平，是根据各年龄组的平均人口数及相应的死亡数计算出来的，${}_nm_X = {}_nD_X / {}_nP_X$。而寿命表中的死亡概率是按某特定人群的年龄别死亡水平，在同时出生的一代人中，X 岁尚存者在今后 n 年内死亡的可能性。

二者的联系：

当年龄组分得较细时，两指标呈下列函数关系：

$$ {}_nq_X = (2n\,{}_nm_X)/(2 + n\,{}_nm_X) $$

或　$q_X = m_X/[1+(1-a_X)m_X]$

其中 a_X 为 X～X+1 岁间死亡者的平均存活年数。0～岁组死亡概率也可以用婴儿死亡率或校正婴儿死亡率来代替。

2．平均年龄是指死者死亡时年龄的算术均数，它取决于年龄别人口构成，两地的平均死亡年龄不能直接进行比较。平均寿命是指 0 岁组预期寿命，是同时出生的一代人按照某年某地的年龄别死亡率水平死亡可预期生存年数。它是各年龄组死亡率的综合反映，不受人口年龄构成的影响，可直接进行比较。

3．在简略寿命表中，$m_{85(+)}$ 表示 85 岁及以上组的年龄组死亡率，即 85 岁及以上组人口在一年内的平均死亡率，它是根据各年龄组的平均人口数计算出来的。而 e_{85} 是指 85 岁的预期寿命。表示 85 岁尚存者预期平均尚能存活的年数（即岁数）。

(四) 计算题

1．解：

（1）求年龄组死亡率（${}_nm_X = \dfrac{{}_nD_X}{{}_nP_X}$），计算结果列入表 16-2 第四栏。

（2）求死亡概率 ${}_nq_X$。计算结果列入第五栏。

其中 q_0 用婴儿组死亡率代 $q_0 = \dfrac{246}{18753} = 0.013118$

最后一个组死亡概率为 1.000000。

（3）尚存人数 l_X 与死亡人数 ${}_nd_X$。首先给定 $l_0 = 100000$，再按前面的计算公式 16-2 及公式 16-3 交替计算。结果列入第六栏和第七栏。

$$ d_0 = l_0 q_0 = 100000 \times 0.013118 = 1311 $$
$$ l_1 = l_0 - d_0 = 100000 - 1311 = 98689 $$
$$ d_1 = l_1 q_1 = 98689 \times 0.004406 = 434 $$
$$ l_2 = l_1 - d_1 = 98689 - 434 = 98255 $$

(4) 求生存人年数 ${}_nL_X$，结果列入第八栏。

本例 $a_0 = 0.145$

$$ L_0 = l_1 + a_0 \times d_0 = 98698 + 0.1450 \times 1311 = 98879 $$
$$ L_{80(+)} = \frac{l_{80}}{m_{80(+)}} = \frac{24212}{0.192551} = 125743 $$

(5) **求生存总人年数**$T_X = \sum {}_nL_X$ **。列入第九栏。**

对${}_nL_X$ **自下而上进行累加**

$$T_{80} = L_{80(+)} = 125743$$

$$T_{75} = L_{75} + T_{80} = 293473$$

(6) **求预期寿命（**$e_X = \frac{T_X}{l_X}$**）。结果列入第十栏。**

$$e_0 = \frac{T_0}{l_0} = \frac{6994553}{100000} = 69.95$$

$$e_1 = \frac{T_1}{l_1} = \frac{6895674}{98689} = 69.87$$

至此，寿命表编制完成，见表 16-2。

表 16-2 1998 **年某市男性居民简略寿命表**

年龄组（岁） $X\sim$ (1)	平均人口数 ${}_nP_X$ (2)	实际死亡人数 ${}_nD_X$ (3)	年龄组死亡率 ${}_nm_X$ (4)	死亡概率 ${}_nq_X$ (5)	尚存人数 l_X (6)	死亡人数 ${}_nd_X$ (7)	生存人年数 ${}_nL_X$ (8)	生存总人年数 T_X (9)	平均预期寿命 e_X (10)
0~	18753	246	0.013118	0.013118	100000	1312	98878	6994553	69.95
1~	54325	60	0.001104	0.004406	98689	434	393888	6895674	69.87
5~	64063	46	0.000718	0.003584	98255	352	490395	6501786	66.17
10~	94683	64	0.000676	0.003374	97903	330	488690	6011391	61.40
15~	114332	90	0.000787	0.003928	97573	383	486907	5522701	56.60
20~	126941	123	0.000969	0.004833	97190	469	484777	5035794	51.81
25~	118930	127	0.001068	0.005325	96721	515	482317	4551017	47.05
30~	91922	104	0.001131	0.005641	96206	542	479675	4068700	42.29
35~	62290	92	0.001477	0.007358	95664	703	476562	3589025	37.52
40~	56806	134	0.002359	0.011725	94961	1113	472022	3112463	32.78
45~	65863	239	0.003629	0.017981	93848	1687	465022	2640441	28.14
50~	54243	346	0.006379	0.031393	92161	2893	453572	2175419	23.60
55~	43355	528	0.012179	0.059093	89268	5275	433152	1721847	19.29
60~	32004	763	0.023841	0.112499	83993	9449	396342	1288695	15.34
65~	24445	972	0.039763	0.180837	74544	13480	339020	892353	11.97
70~	12818	897	0.069980	0.297799	61064	18184	259860	553333	9.06
75~	5813	647	0.111302	0.435368	42880	18668	167730	293473	6.84
80~	2685	517	0.192551	1.000000	24212	24212	125743	125743	5.19

（王仁安　张玉海）

第十七章　　随访资料的生存分析

一、教学大纲要求

（一）掌握内容

1．生存分析基本概念

生存时间、完全数据、截尾数据、死亡率、死亡概率、生存概率、生存率。

2．估计生存率的方法：Kaplan-Meier 法、寿命表法。

（二）熟悉内容

1．生存曲线、半数生存期。

2．生存资料的基本要求。

3．两生存曲线的比较的对数秩检验。

（三）了解内容

Cox 回归模型。

二、教学内容精要

（一）生存分析中的基本概念

1.生存时间（survial time）指观察到的存活时间，如表 11-1 中 t 分别为 360,990,1400,1800 天。生存时间有两种类型：

（1）完全数据（complete　data）指从起点至死亡所经历的时间，即死者的存活时间，如表 11-1 中 360，990，1800 天。

（2）截尾数据（censored　data）由于失访、改变防治方案、研究时间结束时事件尚未发生等情况，使得部分病人不能随访到底，称之为截尾。从起点至截尾所经历的时间，称为截尾数据，如表 11-1 中 1400 天，习惯上记为 1400^{+} 天。

表 11-1　4例鼻咽癌随访记录

患者序号	性别（男=1）	处理组号	开始日期	终止日期	结局（死=1）	存活天数
1	0	1	11/29/80	11/04/85	1	360
2	1	1	06/13/82	06/08/83	1	990
3	1	0	03/02/83	12/31/86	0	1400^{+}
4	0	0	08/04/83	04/10/86	1	1800

2．死亡概率与生存概率

（1）死亡概率（mortality　probability）指死于某时段内的可能性大小，记为 q。年死亡概率的计算公式为 $q=\frac{\text{某年内死亡数}}{\text{某年年初观察例数}}$，若年内有截尾，则分母用校正人口数（校正人口数=年

初人口数-$\frac{1}{2}$截尾例数)。

这里的死亡概率与通常所说的死亡率是有区别的，死亡率的分母常用年平均人口，反映过去一年的死亡频率（年平均水平），而死亡概率则用年初人口，表示往后的一年中死亡机会大小。

（2）生存概率（survival probability）与死亡概率相对应，记为 p，表示在某单位时段开始时存活的个体到该时段结束时仍存活的机会大小。年生存概率的计算公式为 $p=1-q=\frac{\text{某年活满一年人数}}{\text{某年年初人口数}}$，若年内有截尾，也要用校正人口数。

（二）生存率的 Kaplan-Meier 法与寿命表法估计

1．生存率

（1）生存率（survival rate） 指病人经历 t_k 个单位时间后仍存活的概率，记为 $S(t_k)$。若无截尾数据，则

$$S(t_k)=P(T\geq t_k)=\frac{t_k\text{时刻仍存活的例数}}{\text{观察总例数}} \qquad (11\text{-}1)$$

其中 T 为病人的存活时间。如果含有截尾数据，分母就必须分时段校正，故此式一般不能直接应用。

(2) 生存率估计的概率乘法原理

假定病人在各个时段生存的事件独立，生存概率为 p_1, p_2, Λ, p_k，则应用概率乘法得生存率估计的应用公式为

$$S(t_k)=P(T\geq t_k)=p_1 p_2 \Lambda\ p_k \qquad (11\text{-}2)$$

若式中 p_1, p_2, Λ, p_k 用校正人数估计，便可处理截尾数据。

生存概率与生存率在意义上差别很大，前者是单个时段的概率，后者是从 0 至 t_k 多个时段的累积结果。

(3) 生存曲线(survival curve) 指将各个时点的生存率连接在一起的曲线图。

(4) 半数生存期（median survival time） 表示恰好有 50%的个体可活这么长时间。

2．生存率的估计方法

（1）乘积极限法（product-limit method） 直接用概率乘法原理估计生存率（不分组），由 Kaplan-Meier 于 1958 年提出，因而又称 Kaplan-Meier 法。这是一种非参数法，主要用于小样本，也适用于大样本。其生存曲线是左连续的阶梯形曲线。

（2）寿命表法（life-table method） 当样本例数足够多时，乘积极限法可按时间分组计算，这就是寿命表法，实际上是乘积极限法的一种近似。其生存曲线呈折线形。

（三）两样本生存曲线的比较——对数秩检验

对数秩检验（log-rank test）用于两样本生存曲线的比较，其零假设为两总体生存曲线相同，基本思想是如果零假设成立，根据不同日期两种处理的期初人数和死亡人数，计算各种处理在各个时期的理论死亡数。若零假设成立，则实际死亡数与理论死亡数不会相差太大，否则应认为零假设不可能成立，两条生存率曲线差异有统计学意义。

对数秩检验统计量（近似法）为：

$$c^2=\sum_{k=1}^{m}\frac{(A_k-T_k)^2}{T_k} \quad ，u=m-1 \qquad (11\text{-}3)$$

其中 A_k 和 T_k 分别是第 k 组死亡的实际数和理论期望数。在 H_0 成立的条件下，统计量 c^2 服从自由度为 $m-1$ 的 c^2 分布，m 为组数，据 c^2 作出是否拒绝 H_0 的决定。

（四）Cox 回归模型

Cox 回归是生存分析中最重要的方法之一，其优点是适用范围很广和便于做多因素分析。

Cox回归假定病人的风险函数为

$$h(t)=h_0(t)\exp(b_1X_1+b_2X_2+\Lambda+b_pX_p) \qquad (11\text{-}4)$$

其中$h(t)$为风险函数，又称风险率或瞬间死亡率，$h_0(t)$为基准风险函数，是与时间有关的任意函数，X,b分别是观察变量及其回归系数。英国统计学家Cox D R 提出了参数b_i的估计和检验方法，故称为Cox回归。

三、典型试题分析

（一）单项选择题

1．生存分析的效应变量是（　　）。

A．正态的和方差齐性的　　B．生存时间和结局变量

C．生存时间　　D．结局变量

答案：B

[评析]　本题考点：生存分析的概念

生存分析是将事件的结果和出现这一结果所经历的时间,结合起来分析的一种统计分析方法，所以它的应变量有两个，即生存时间和结局。

2．随访资料做生存分析的条件为(　　)。

A．有一定的例数　　B．有一定的死亡数

C．死亡比例不能过小　　D．自变量取值不随时间变化

答案：B

[评析] 本题考点：生存资料的基本要求

生存资料的基本要求为：①样本由随机抽样方法获得，并有足够数量；②死亡例数不能太少（≧30)；③截尾比例不能太大；④生存时间尽可能精确到天数；⑤缺项要尽量补齐。所以最佳答案应选B。

3．Cox回归风险率(　　)。

A．等于一个常数　　B．服从某种分布规律

C．等于基准函数乘上一个比例因子　　D．适用于任意肿瘤资料

答案：C

[评析]　本题考点：Cox回归模型的特点及应用

首先，用于 Cox 回归模型分析的资料必须满足生存资料的基本要求，因此任意肿瘤资料不一定满足此要求，排除 D。Cox 回归风险函数中因$h_0(t)$未定义，所以不知道风险在病人与病人之间的差别和风险随时间变化的具体分布，排除 A，B。所以正确答案为 C，从风险回归函数的定义式也可看出。

4．采用 log-rank 检验分析肺癌发病资料，其中吸烟、慢性支气管炎 2 个因素都有统计学意义，由此可认为(　　)：

A．吸烟与肺癌有因果联系　　B．慢性支气管炎与肺癌有因果联系

C．2 个因素与肺癌有因果联系　　D．以上都不对

答案：D

[评析]　本题考点：模型中的变量如何选择取舍

选入模型的变量是统计学上的有关变量，不一定都与肺癌有因果关系，其中某些可能只

有伴随关系而已；未选入模型的变量不一定全是无关变量，要考虑是否模型内的某些变量代替了它的作用，或因例数不够，或实验中对该因素进行了控制而引起的。所以正确答案选D。

5．根据表 11-1 中的存活时间，试用 Kaplan-Meier 法估计生存曲线。

[评析]　本题是考察对乘积极限法的应用情况，此法应用普遍，应熟练掌握。具体解法见表 11-2。

表 11-2　乘积极限法估计生存率计算表

序号 k	存活时间(天) t	t 时刻期初例数 n	t 时刻死亡数 d	死亡概率 $q=d/n$	生存概率 $p=1-q$	k 年生存率 $S(t_k)$	生存率标准误 $SE(S(t_k))$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	360	4	1	1/4	3/4	(3/4)=0.75	0.2165
2	990	3	1	1/3	2/3	(3/4)(2/3)=0.50	0.2500
3	1800	1	1	1/1	0/1	(3/4)(2/3)(0/1)=0.00	0

参照表 11-2，计算步骤为：

1．列出序号： k =1，2，...**（第 1 列）；**

2．死亡时间排队：将存活时间 t **（完全数据）从小到大顺序排列，重复数据只列一次，截尾数据（如** 1400^+ **) 不列入（第 2 列）；**

3．求出 t **时刻期初例数** n **：即存活时间大于或等于** t **的例数（含死者）（第 3 列）；**

4．列出 t **时刻的死亡数** d **：即死亡时间为** t **的例数（第 4 列）；**

5．求出 t **时刻的死亡概率：（第 5 列）；**

6．求出 t **时刻的生存概率：（第 6 列）；**

7．分别计算生存率及其标准误；（第 7、8 列）；

8．绘制生存曲线。

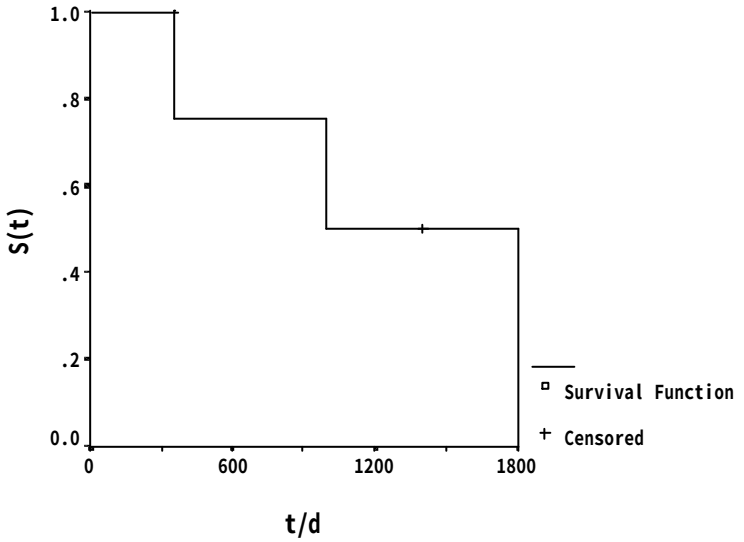


图11.1　乘积极限法生存曲线及其半数生存期

四、习　题

（一）名词解释

1.生存分析　2.生存时间　3.完全数据　4.截尾数据　5.死亡率

6.死亡概率　7.生存概率　8.生存率

（二）单项选择题

1．Cox回归的自变量(　　)。

A．必须服从正态分布和方差齐性　　B．必须是计量资料

C．可以是计量资料或分类资料　　D．无任何条件

2．生存分析中的生存时间为(　　)。

A．出院至失访的时间　　B．手术至失访的时间

C．观察开始至终止的时间　　D．观察开始至失访的时间

3．关于膀胱癌化疗的随访资料做生存分析，可当作截尾值处理的是(　　)：

A．死于膀胱癌　　B．死于意外死亡

C．死于其它肿瘤　　D．b,c都是

（三）简答题

1．在肿瘤预后分析中，死于非肿瘤患者的数据怎样处理？

2．生存分析可用于发病资料的分析吗？请举例说明。

3．生存时间能计算平均数、标准差吗？

4．Cox回归可估计参数，故属于参数方法？

（四）计算题

1．表11-3第2-4列是296例肝癌患者的生存数据，试作生存分析并绘图示之。

2．某院用甲、乙两疗法组治疗急性黄疸性肝炎，随访十年得资料如下：

甲疗法组　12，25，50^+，68，70，79^+，83^+，91^+，114^+，114^+，

乙疗法组　1，1，9，17，21，25，37，38，58，72^+，73

比较两疗法的生存期（月）有无差别。

五、习题答题要点

（一）名词解释

1．**生存分析：生存分析**（survival　analysis）**是将事件的结果和出现这一结果所经历的时间，结合起来分析的一种统计分析方法，它不仅可以从事件结局的好坏，如疾病的治愈（成功）和死亡（失败），而且可以从事件的持续时间，如某病经治疗后存活的时间长短进行分析比较，因而能够更全面、更精确地反映该治疗的效果。**

2．**生存时间：生存时间**（survival　time）**指观察到的存活时间。**

3．**完全数据：完全数据**（complete　data）**指从起点至死亡所经历的时间，即死者的存活时间。**

4．截尾数据：由于失访、改变防治方案、研究时间结束时事件尚未发生等情况，使得部分病人不能随访到底，称之为截尾。从起点至截尾所经历的时间，称为截尾数据（censored data）。

5．死亡率：某年内死亡例数与年中观察例数之比称为死亡率（mortality rate）。

6． 死亡概率：死亡概率（mortality probability）是指某年内死亡例数与年初观察例数之比，若年内有截尾，分母用校正人口数。

7．生存率：生存率（survival rate）指病人经历 t_k 个单位时间后仍存活的概率，即 t_k 时刻仍存活的例数与观察总例数之比。

8．生存概率：生存概率（survival probability）表示在某单位时段开始时存活的个体到该时段结束时仍存活的机会大小，它是某年活满一年人数与年初观察例数之比，若年内有截尾，分母用校正人口数。

（二）单项选择题

1.C 2.C 3.D

（三）简答题

1．当作截尾数据处理。

2．可用于慢性病的发病资料分析。

3．如果此资料所包含的数据都是完全数据，可以计算均数和标准差（但可能因资料非正态而没有实际意义），若数据中包含截尾数据，则不可以计算均数和标准差。

4．属于半参数模型（因 $h_0(t)$ 未定义）。

（四） 计算题

1．参照表 11-3，列表计算。

表 11-3 寿命表法估计 296 例肝癌患者生存率计算表

序号 k	存活时间（月） t	期内死亡人数 d	期内截尾人数 c	期初观察人数 n_0	校正年初人数 $n_c = n_0 - c/2$	死亡概率 $q = d/n$	生存概率 $p = 1-q$	k 年生存率 $S(t_k)$	生存率标准误 SE($S(t_k)$)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	0~	94	10	296	291.0	0.3230	0.6770	0.6770	0.0274
2	1~	74	15	192	184.5	0.4011	0.5989	0.4055	0.0294
3	2~	22	10	103	98.0	0.2245	0.7755	0.3144	0.0285
4	3~	22	6	71	68.0	0.3235	0.6765	0.2127	0.0263
5	4~	5	5	43	40.5	0.1235	0.8765	0.1864	0.0255
6	5~	6	6	33	30.0	0.2000	0.8000	0.1492	0.0245
7	6~	4	1	21	20.5	0.1951	0.8049	0.1201	0.0237
8	7~	2	1	16	15.5	0.1290	0.8710	0.1046	0.0230
9	8~	3	2	13	12.0	0.2500	0.7500	0.0784	0.0217
10	9~	2	0	8	8.0	0.2500	0.7500	0.0588	0.0202

11	10~	2	2	6	5.0	0.4000	0.6000	0.0353	0.0177
12	11~	2	2	2	2.0	1.0000	0.0000	0.0000	0.0000

计算方法和步骤为：

（1）列出序号： k =1，2，...**（第** 1 **列）；**

（2）求校正期初人数： $n_c = n_0 - c/2$ **（第** 6 **列）；**

（3）计算死亡概率： $q = d/n$ **（第** 7 **列）；**

（4）计算生存概率： $p = 1 - q$ **（第** 8 **列）；**

（5）计算生存率及其标准误：（第 9、10 **列）。**

（6）绘制生存曲线。

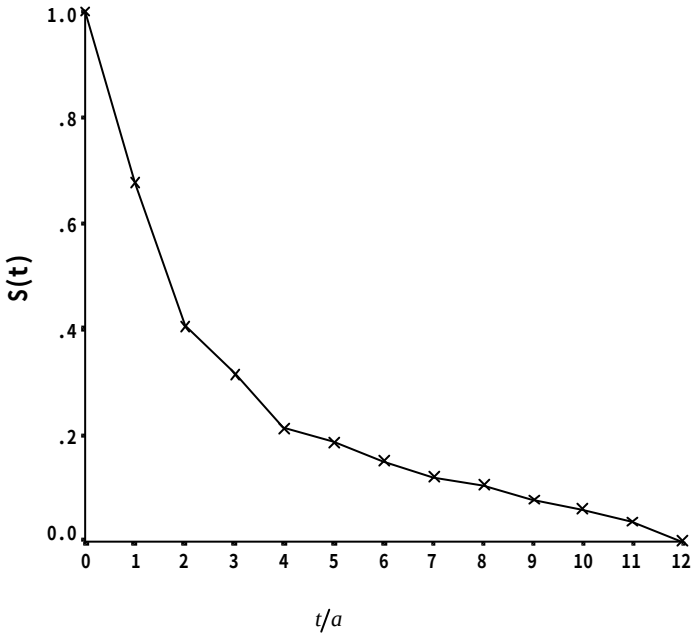


图 11-2　　**寿命表法生存曲线**

2．**该题用** log-rank **检验，其计算步骤为：**

（1）检验假设： H_0 **：两总体生存率曲线相同；** H_1 **：两总体生存率曲线不同；** $a = 0.05$ **。**

（2）计算出 A_0 =4，A_1 =10，T_0 =8.6694，T_1 =5.3306.

（3）计算检验统计量

$$c^2 = \sum_{k=1}^{m} \frac{(A_k - T_k)^2}{T_k} = \frac{(4-8.6694)^2}{8.6694} + \frac{(10-5.3306)^2}{5.3306} = 6.6052$$

（4）确定 P **值作结论：据自由度为** 1 **的** c^2 **分布查表得** P <0.05**，按** $a = 0.05$ **水准拒绝** H_0 **，接受** H_1 **，故可认为两总体生存率曲线不同，甲组疗法生存期长。**

（骆福添　杜晓晗）

第二章　常用综合评价方法

一 、教学大纲要求

（一）掌握内容

综合评价的意义及一般步骤。

（二）熟悉内容

评价指标的筛选及权重的估计。

（三）了解内容

1．综合评分法。

2．综合指数法。

3．层次分析法。

4．Topsis 法。

二、教学内容精要

（一）评价与综合评价

评价：通过对照某些标准来判断观测结果，并赋予这些结果以一定的意义和价值的过程称为评价（evaluation）。

综合评价：根据一个复杂系统同时受到多种因素影响的特点，在综合考察多个有关因素时，依据多个有关指标对复杂系统进行总评价的方法称为综合评价（synthetical evaluation）。

（二）综合评价的几种分类

1．根据评价手段：定量评价(quantitative evaluation)、定性评价(qualitive evaluation)。

2．根据评价领域：临床评价(clinical evaluation)、卫生评价(health evaluation)和管理评价(administrative evaluation)。

临床评价包括诊断性试验和方法评价、疗效评价和预后及转归评价。

卫生评价包括环境评价、营养评价、生长发育评价和疾病防治效果评价。

管理评价包括宏观管理评价和微观管理评价。

3．根据评价方式：预评价(pre-event evaluation)、中期评价(medial evaluation)和终结评价(after-event evaluation)。

（三）综合评价的一般步骤

1．根据评价目的选择恰当的评价指标(index)；

2．根据评价目的，确定诸评价指标在对某事物评价中的相对重要性，或各指标的权重(weight)；

3．合理确定各单个指标的评级等级(evaluation grade)及其界限；

4．根据评价目的，数据特征，选择适当的综合评价方法，并根据已掌握的历史资料，建立综合评价模型(evaluation model)；

5．确定多指标综合评价的等级数量界限，在对同类事物综合评价的应用实践中，对选用的评价模型进行考察，并不断修改补充，使之具有一定的科学性、实用性与先进性，然后推广应用。

（四）评价指标的筛选

筛选评价指标主要依据专业知识，即根据有关的专业理论和实践，来分析各评价指标对结果的影响，挑选那些代表性、确定性好，有一定区别能力又互相独立的指标组成评价指标体系。

系统分析法(system's analysis method)和文献资料分析优选法是常用的评价指标筛选法。

为保证筛选指标的客观性，对于指标的初选可采用假设检验、多元回归、逐步回归和指标聚类等方法辅助筛选。

在实际工作中，往往综合使用多种方法进行指标筛选，在获得较为满意的专业解释的基础上，优先考虑那些被多种方法同时选入的指标。

（五）评价指标的权重估计

用于确定指标权重的方法主要有主观定权法和客观定权法。其中，主观定权法包括专家评分法(specialist-scored method)、成对比较法、Satty 权重法；客观定权法包括模糊定权法、秩和比法、熵权法和相关系数法。

（六）几种综合评价方法

1．综合评分法(synthetical scored method)：建立在专家评价法基础上，根据评价目的及评价对象的特征选定必要的评价指标，逐个指标订出等级，每个等级的标准用分值表示，然后以恰当的方式确定各评价指标的权数，并选定累积总分的方案以及综合评价等级的总分值范围，以此为准则，对评级对象进行分析和评价，以决定优劣取舍的综合评价方法。

2．综合指数法(synthetical index method)：利用综合指数的计算形式，定量的对某现象进行综合评价的方法。

3．层次分析法(analytic hierarchy process)：用系统分析的方法，对评价对象依评价目的所确定的总评价目标进行连续性分解，得到各级（各层）评价目标，并以最下层作为衡量目标达到程度的评价指标。然后依据这些指标计算出一综合评分指数对评价对象的总评价目标进行评价，依其大小来确定评价对象的优劣等级。

4．Topsis法：系统工程中有限方案多目标决策分析的一种常用方法。是基于归一化后的原始数据矩阵，找出有限方案中的最优方案和最劣方案（分别用最优向量和最劣向量表示），然后分别计算诸评价对象与最优方案和最劣方案的距离，获得各评价对象与最优方案的相对接近程度，以此作为评价优劣的依据。

三、典型试题分析

某医院 1998 年 11 项指标资料见表 18-1 和表 18-2，试采用综合指数法计算各月综合指数。

表 18-1　11 项指标分类

指标类型	序号	指标名称

动态指标	1	出院病人数（人）
医疗质量	2	治疗有效率（%）
	3	病死率（%）
	4	无菌手术感染数（人）
床位利用	5	平均住院日（天）
	6	床位周转率（%）
	7	病床工作日（天）
	8	病床使用率（%）
诊断水平	9	门诊住院诊断符合率（%）
	10	出入院诊断符合率（%）
护理服务质量	11	陪住率（%）

表 18-2　某医院 1998 年各月 11 项指标实际值

月	各指标实际值										
	1	2	3	4	5	6	7	8	9	10	11
1	650	90.8	3.08	3.00	20.6	1.41	28.7	92.6	99.3	100	18.0
2	560	91.1	3.04	4.00	21.6	1.24	28.7	92.7	98.6	100	17.6
3	609	91.7	1.97	4.00	20.5	1.33	27.3	97.6	98.0	99	17.0
4	587	92.7	2.39	3.00	25.6	1.25	30.0	96.9	98.3	96	17.1
5	651	88.0	4.30	4.00	23.3	1.30	28.3	94.5	97.3	97	18.0
6	601	89.7	2.50	10.00	19.8	1.30	29.3	94.6	97.9	96	17.0
7	584	90.0	2.91	5.00	26.3	1.30	28.0	93.2	96.9	97	18.0
8	620	90.7	2.90	2.00	22.0	1.37	28.7	92.5	97.9	96	19.0
9	626	90.2	2.24	4.00	22.0	1.37	29.2	94.3	98.3	98	18.0
10	604	91.9	2.96	5.00	20.6	1.34	27.9	93.1	99.1	99	18.5
11	653	90.5	3.37	2.00	19.5	1.44	29.4	94.8	99.5	99	21.0
12	599	90.8	3.64	5.00	23.5	1.29	28.7	95.8	89.1	99	18.8
平均值	612	90.7	2.94	4.25	22.1	1.33	28.7	94.4	97.5	98	18.2

［评析］ 11 项指标中 3、4、5、11 号指标为反向指标，其它均为正向指标。由公式（18-1）和公式（18-2）可计算出各指标的个体指数，计算结果见表 18-3。

$$y=\frac{X}{M}\quad\text{（高优指标或正指标）}\qquad(18\text{-}1)$$

$$y=\frac{M}{X}\quad\text{（低优指标或负指标）}\qquad(18\text{-}2)$$

如公式（18-1）和公式（18-2）所示，个体指数是某指标观测值和标准值的比值。式中 X 为某指标的观测值；M 为某指标的标准值、参考值、平均值、期望值等。

表 18-3　某医院 1998 年各月 11 项指标的个体指数

月	各指标的个体指数										
	1	2	3	4	5	6	7	8	9	10	11
1	1.06	1.00	0.96	1.42	1.07	1.06	1.00	0.98	1.02	1.02	1.01
2	0.91	1.00	0.97	1.06	1.02	0.93	1.00	0.98	1.01	1.02	1.03
3	1.00	1.01	1.49	1.06	1.08	1.00	0.95	1.03	1.00	1.01	1.07
4	0.96	1.02	1.23	1.42	0.86	0.94	1.05	1.03	1.01	0.98	1.06
5	1.06	0.97	0.68	1.06	0.95	0.98	0.99	1.00	1.00	0.99	1.01
6	0.98	0.99	1.18	0.42	1.12	0.98	1.02	1.00	1.00	0.98	1.07
7	0.95	0.99	1.01	0.85	0.84	0.98	0.98	0.99	0.99	0.99	1.01
8	1.01	1.00	1.01	2.13	1.00	1.03	1.00	0.98	1.00	0.98	0.95
9	1.02	0.99	1.31	1.06	1.00	1.03	1.02	1.00	1.01	1.00	1.01
10	0.99	1.01	0.99	0.85	1.07	1.01	0.97	0.99	1.02	1.01	0.98
11	1.07	1.00	0.87	2.12	1.13	1.08	1.02	1.00	1.02	1.01	0.87
12	0.98	1.00	0.81	0.85	0.94	0.97	1.00	1.01	0.91	1.01	0.97

按同类指数相乘，异类相加的方法进行指数综合。由公式（18 - 3）计算综合指数。

$$I=\sum_{i=1}^{m}\prod_{j=1}^{n}y_{ij} \qquad (18\text{-}3)$$

例如计算 1、2 月份的综合指数为：

$$I_1=1.06+1.00\times0.96\times1.42+1.07\times1.06\times1.00\times0.98+1.02\times1.02+1.01$$
$$=5.5851$$

$$I_2=0.91+1.00\times0.97\times1.06+1.02\times0.93\times1.00\times0.98+1.01\times1.02+1.03$$
$$=4.9509$$

其余各月计算以次类推，计算结果见表 18-4。

表 18-4　某医院 1998 年各月综合指数

月份	1	2	3	4	5	6	7	8	9	10	11	12
指数	5.5851	4.9509	5.7462	5.6642	4.6833	4.6476	4.5929	6.1251	5.4824	4.8934	6.0795	4.4838

四、习　题

（一）单项选择题：

1．下列那项评价方法属于按评价手段的分类

A．定性评价　　B．卫生评价

C．管理评价　　D．中期评价

2．使用专家评分法进行评价指标的估计时，常用哪两种指标来估计权重分配的相对合理性

A．擅长系数和确定系数　　B．擅长系数和一致性系数

C．相关系数和确定系数　　D．相关系数和一致性系数

3．医院工作质量指标通常由三层子指标构成，以知第一层的权重为 0.6370，第二层权重为 0.2970，第三层权重为 1.0。由 Saaty 法提供的评价指标组合权重方法可知第三层的组合权重为

A．1.9340　　B．0.9340　　C．1.7636　　D．0.1892

4．以下哪一种综合评价方法是建立在专家评价法的基础上

A．综合指数法　　B．层次分析法

C．综合评分法　　D．Topsis 法

5．在利用综合指数法评价时，综合指数能定量地反映几个指标的综合平均变动程度，表达式为：$I=\frac{1}{n}\sum_{1}^{m}y$，其中

A．m 为分组数　　B．n 为指标数

C．y 为个体指标　　D．以上均正确

（二）计算题

试根据表 18 - 5 数据，采用 Topsis 法对某市人民医院 1995~1997 年的医疗质量进行综合评价。

表 18 - 5　某市人民医院 1995~1997 年的医疗质量

年度	床位周转次数	床位周转率（%）	平均住院日	出入院诊断符合率（%）	手术前后诊断符合率（%）	三日确诊率（%）	治愈好转率（%）	病死率（%）	危重病人抢救成功率（%）	院内感染率（%）
1995	20.97	113.81	18.73	99.42	99.80	97.28	96.08	2.57	94.53	4.60
1996	21.41	116.12	18.39	99.32	99.14	97.00	95.65	2.72	95.32	5.99
1997	19.13	102.85	17.44	99.49	99.11	96.20	96.50	2.02	96.22	4.79

五、习题答题要点

（一）单项选择题

1.A　2.B　3.D　4.C　5.C

（二）计算题

对原指标中的平均住院日、病死率、院内感染率三个低优指标进行转化，其中平均住院日采用倒数转化，病死率、院内感染率采用差值转化。转化后数据见表 18-6。

表 18-6　转化指标值

年度	床位周转次数	床位周转率（%）	平均住院日	出入院诊断符合率（%）	手术前后诊断符合率（%）	三日确诊率（%）	治愈好转率（%）	病死率（%）	危重病人抢救成功率（%）	院内感染率（%）
1995	20.97	113.81	5.34	99.42	99.80	97.28	96.08	97.43	94.53	95.40
1996	21.41	116.12	5.44	99.32	99.14	97.00	95.65	97.28	95.32	94.01
1997	19.13	102.85	5.73	99.49	99.11	96.20	96.50	97.98	96.22	95.21

根据表 18-6 数据，利用公式（18-4）进行归一化处理，得归一化矩阵值，如表 18-7。

$$Z_{ij}=\frac{X_{ij}}{\sqrt{\sum_{i=1}^{n}\left(X_{ij}\right)^2}} \tag{18-4}$$

例如计算 1995 年床位周转次数归一化值，由公式（18-4）得：

$$Z_{11}=\frac{20.97}{\sqrt{20.97^2+21.41^2+19.13^2}}=0.509$$

其余归一化数值以此类推。

表 18-7　归一化矩阵值

年度	床位周转次数	床位周转率	平均住院日	出入院诊断符合率	手术前后诊断符合率	三日确诊率	治愈好转率	病死率	危重病人抢救成功率	院内感染率
1995	0.590	0.592	0.560	0.577	0.580	0.580	0.577	0.577	0.572	0.581
1996	0.602	0.604	0.570	0.577	0.576	0.578	0.575	0.576	0.577	0.572
1997	0.538	0.535	0.601	0.578	0.576	0.574	0.580	0.580	0.583	0.579

由公式（18-5）和公式（18-6）得最优和最劣方案

最优方案　$Z^+=\left(a_{i1\max},a_{i2\max},\Lambda\ ,a_{im\max}\right)$　（18-5）

最劣方案　$Z^-=\left(a_{i1\min},a_{i2\min},\Lambda\ ,a_{im\min}\right)$　（18-6）

$$Z^+=(0.602,0.604,0.601,0.578,0.580,0.580,0.580,0.580,0.583,0.581)$$

$$Z^-=(0.538,0.535,0.560,0.577,0.576,0.574,0.575,0.576,0.572,0.572)$$

由公式（18 - 7）和公式（18 - 8）计算各年度 D^+ 和 D^-，见表 18 – 8。

$$D_i^+=\sqrt{\sum_{i=1}^{m}\left(a_{ij\max}-a_{ij}\right)^2} \tag{18-7}$$

$$D_i^-=\sqrt{\sum_{i=1}^{m}\left(a_{ij\min}-a_{ij}\right)^2} \tag{18-8}$$

例如计算 1997 年 D^+ 和 D^-：

$$D^+=\sqrt{(0.602-0.538)^2+(0.604-0.535)^2+\Lambda\ +(0.581-0.579)^2}$$
$$=0.094$$
$$D^-=\sqrt{(0.538-0.538)^2+(0.535-0.535)^2+\Lambda\ +(0.572-0.579)^2}$$
$$=0.044$$

其余各年以次类推。

由公式（18-9）计算各年度 C_i，见表 18 - 8。

$$C_i=\frac{D_i^-}{D_i^++D_i^-} \tag{18-9}$$

例如计算 1997 年 C_i：

$C_i=\dfrac{0.044}{0.094+0.044}=0.319$，其余各年以次类推。

表 18-8　不同年度指标值与最优值的相对接近程度及排序结果

年份	D^+	D^-	C_i	排序结果
1995	0.045	0.078	0.634	2
1996	0.034	0.095	0.736	1
1997	0.094	0.044	0.319	3

由表 18-8 **的排序结果可知** 1996 **年医疗质量最好。**

（孙振球　潘峰）