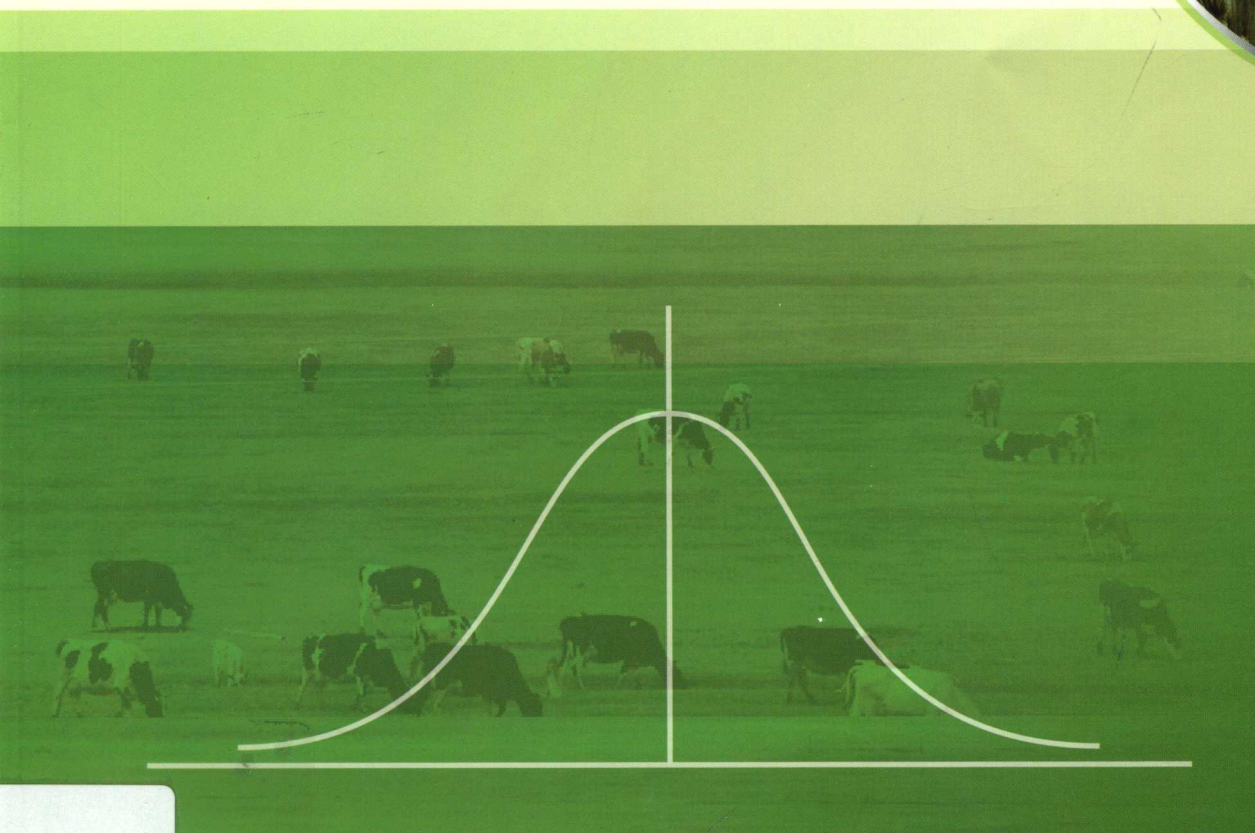


“十二五”普通高等教育本科国家级规划教材
《生物统计学》（第五版）立体化配套教材

生物统计学学习指导

（第二版）

李春喜 姜丽娜 邵云 张黛静 编著



科学出版社

Q-332
2009.2

阅 览

“十二五”普通高等教育本科国家级规划教材
《生物统计学》(第五版)立体化配套教材

生物统计学学习指导

(第二版)

李春喜 姜丽娜 邵云 张黛静 编著



科学出版社

北京

内 容 简 介

本书是“十二五”普通高等教育本科国家级规划教材《生物统计学》(第五版)的立体化项目之一。本书旨在为《生物统计学》的学习提供概要性总结、资料扩充、难点解析,通过增加具体实例和对习题的解答,帮助学生进一步理解和掌握基本概念、基本内容和基本方法。全书共13章,其内容编排与教材各章内容相对应。每章内容包括目的要求、内容提要、难点评析、例题解析、习题解答、自我测验6部分。书后附有自我测验答案。

本书可作为综合性大学、师范院校生物类及相关专业本科生学习《生物统计学》的配套学习辅导书,也可作为生命科学、生物工程、农业科学、林业科学、医学、畜牧兽医、水产科学等专业的科研工作者、教师和研究生的参考书。

图书在版编目(CIP)数据

生物统计学学习指导/李春喜等编著.—2版.—北京:科学出版社,2013
“十二五”普通高等教育本科国家级规划教材.《生物统计学》(第五版)立体化
配套教材

ISBN 978-7-03-037629-9

I.①生… II.①李…②姜…③邵… III.①生物统计-高等学校-教学参考资料
IV.①Q-332

中国版本图书馆CIP数据核字(2013)第116335号

责任编辑:丛楠 贺窑青/责任校对:朱光兰

责任印制:阎磊/封面设计:迷底书装

科学出版社出版

北京东黄城根北街16号

邮政编码:100717

<http://www.sciencep.com>

源海印刷有限责任公司印刷

科学出版社发行 各地新华书店经销

*

2008年7月第一版 开本:787×1092 1/16

2013年6月第二版 印张:15

2013年6月第四次印刷 字数:340 000

定价:29.00元

(如有印装质量问题,我社负责调换)

第二版前言

生物统计学的学习离不开举一反三。学好、用好生物统计原理与方法,不但要对其基本概念、基本内容有较熟悉的理解和掌握,也要通过例题学习了解不同统计问题的解题思路和解题方法,更要通过习题练习来熟练掌握这些方法。为了配合《生物统计学》(第五版)的出版,便于教师教好、学生学好生物统计学,帮助学生更好、更快地掌握生物统计学的基本理论、试验设计和统计分析方法,我们组织编写了这本《生物统计学学习指导》。

本书共 13 章,在内容体系安排上与《生物统计学》(第五版)保持一致。每章包括目的要求、内容提要、难点评析、例题解析、习题解答、自我测验 6 部分。目的要求部分提出本章要达到的基本要求。内容提要部分概要地介绍本章的主要知识点和难点、关键点。难点评析部分对本章的疑难问题进行较细致的剖析,适当扩充了部分内容,对重要问题的解题思路、解题方法及注意事项作了介绍。例题解析部分是在《生物统计学》(第五版)例题的基础上,重点选取部分有代表性的例子对其解题过程进行系统分析、计算和评述。习题解答部分对教材每章后所附思考练习题一一进行详细解答。自我测验部分则结合《生物统计学》各类考题形式,设计部分题目,主要包括填空、判断、名词解释、单项选择和计算 5 种类型,供读者选择练习。书后附有自我测验答案,供学生参考。

在本书编写和出版过程中,得到了科学出版社、河南师范大学的大力支持,特别是科学出版社王国栋先生、丛楠女士对本书的编写给予了多方面的指导和帮助,在此一并感谢。

由于作者对编写学习指导书的经验有限,掌握的素材也不够多,本书在内容设计与编排、解题方法和技巧等方面会存在许多不足之处。殷切希望广大读者对书中的疏漏和不妥及时给予批评指正,以便本书再版时进一步完善。

李春喜 姜丽娜 邵 云 张黛静

2013 年 3 月于河南师范大学

第一版前言

生物统计学是一门实用性很强的工具性课程。学习生物统计学需要举一反三,既要
对生物统计学的基本概念、基本内容有一定的理解和掌握,也要通过例题学习来了解不同
统计问题的解题思路和解题方法,更要通过习题练习来熟练掌握这些方法。因此,编写一
本与《生物统计学》教材配套的学习指导书就显得十分必要。多年来,作者编写的《生物统
计学》得到了广大读者的厚爱,其使用范围也在不断扩大。但由于在教材编写时受课程教
学的限制,内容体系只能涉及基本的统计问题和部分扩展性知识,用于介绍和解析各种统
计方法的例题也只能选择少部分经典实例,这就不可避免地会使一些问题得不到细致分
析,部分内容的叙述和公式推导也不够深入。对习题部分,也只是给出简单的参考答案,
而没有细致的解题过程。因此,不断有读者提出上述问题,询问其相关解决方法。为了有
效解决上述问题,更好地配合生物统计学教学,我们结合《生物统计学》(第四版),组织编
写了这本《生物统计学学习指导》。

本书在内容体系安排上与《生物统计学》(第四版)保持一致,共 14 章。内容包括目的
的要求、内容提要、难点评析、例题解析、习题解答和自我测验等 6 部分。书后附有自我测验
答案。目的要求部分提出了本章要达到的基本要求;内容提要部分概要地介绍了本章的
主要知识点和难点、关键点;难点评析部分是对本章的疑难问题进行较细致的剖析,适当
扩充了部分内容,对重要问题的解题思路、解题方法以及注意事项作了介绍;例题解析部
分是在教材例题的基础上,重点选取部分代表性的例子对其解题过程进行了系统分析、计
算和评述;习题解答部分对教材每章后所附思考练习题一一进行了详细解答。自我测验
部分则是结合《生物统计学》各类考题形式,设计了部分题目,主要包括填空、判断、名词解
释、单项选择和计算等 5 种类型,供读者练习。书后附有自我测验答案,供参考。

本书在编写和出版过程中,得到了科学出版社甄文全先生、周辉先生、王国栋先生和
河南师范大学教务处、生命科学学院的大力支持,在此一并表示感谢。

由于作者在编写学习指导书时缺乏经验,掌握的素材也不够多,在内容设计与编排、
解题方法和技巧等方面会存在许多不足之处。殷切希望广大读者对书中的疏漏和不妥之
处及时给予批评指正,以便本书再版时进一步完善。

李春喜 姜丽娜 邵云

2008 年 3 月于河南师范大学

目 录

第二版前言	
第一版前言	
第一章 概论	1
目的要求	1
内容提要	1
难点评析	1
习题解答	4
自我测验	5
第二章 试验资料整理与特征数计算	7
目的要求	7
内容提要	7
难点评析	8
例题解析	11
习题解答	13
自我测验	18
第三章 概率与概率分布	21
目的要求	21
内容提要	21
难点评析	22
例题解析	24
习题解答	26
自我测验	30
第四章 统计推断	32
目的要求	32
内容提要	32
难点评析	33
例题解析	35
习题解答	39
自我测验	45
第五章 χ^2 检验	47
目的要求	47

内容提要	47
难点评析	47
例题解析	49
习题解答	53
自我测验	57
第六章 方差分析	59
目的要求	59
内容提要	59
难点评析	60
例题解析	62
习题解答	66
自我测验	79
第七章 直线回归与相关分析	82
目的要求	82
内容提要	82
难点解析	83
例题解析	89
习题解答	97
自我测验	102
第八章 可直线化的非线性回归分析	105
目的要求	105
内容提要	105
难点评析	105
例题解析	107
习题解答	115
自我测验	123
第九章 试验设计及其统计分析	124
目的要求	124
内容提要	124
难点评析	125
例题解析	130
习题解答	142
自我测验	157
第十章 协方差分析	160
目的要求	160
内容提要	160

难点评析	160
例题解析	163
习题解答	171
自我测验	180
第十一章 多元线性回归与多元相关分析	182
目的要求	182
内容提要	182
难点评析	182
例题解析	183
习题解答	188
自我测验	195
第十二章 逐步回归与通径分析	197
目的要求	197
内容提要	197
难点评析	197
例题解析	198
习题解答	204
自我测验	210
第十三章 多项式回归分析	212
目的要求	212
内容提要	212
难点评析	212
例题解析	213
习题解答	218
自我测验	223
自我测验答案	225

数据资料,以便对数据的基本特征有清晰、直观的了解。分析资料是针对要研究的问题,通过对数据的深入分析,从数据资料中获取所需有关信息的过程。解释资料是在分析结果的基础上对所研究的问题作出统计推断。综上所述,统计学是与数据密切相关的科学,因而可将统计学看成是应用数学的一个分支。

2. 生物统计学是生物科学研究的基本工具

生物统计学是用数理统计的原理和方法来分析及解释生物界各种现象和试验调查资料的一门学科。生物统计学是生物科学研究和应用必不可少的基本工具,这是由生物现象的基本特征决定的,生物现象有如下基本特征。

(1) 变异性。遗传和变异是生物的两大基本现象,在任何两个生物个体之间均存在差异。

(2) 不确定性(随机性)。生物个体之间的差异往往是由很多偶然因素造成的,因而而是随机的,不能准确预测。

(3) 复杂性。造成生物变异性的因素通常有很多,既有遗传方面的,也有环境方面的。从遗传方面来说,生物个体所携带的基因系统是复杂的,从上代传递给下代的方式也是多种多样的,人们至今对生物的遗传机制也还没有一个完全清楚的认识;从环境方面说,生物的很多性状都受各种环境因素的影响,其中有的是可以人为控制的环境因素,但更多的是不可控制的、对不同个体产生随机性影响的环境因素。

生物现象的这些特点决定了不能通过描述性的定性科学或决定性的数量来解决生物学领域中的众多问题。只能通过对一定数量的数据资料进行分析才能得到答案,而只有生物统计学才能告诉我们如何通过科学的调查或试验获得高质量的数据,如何对所获得的数据进行分析,如何根据分析的结果得出尽可能可靠的结论。

3. 正确理解总体、个体和样本的概念

总体是研究对象的全体,是具有相同性质的个体所组成的集合;个体是组成总体的基本单元;样本是由总体中抽出的若干个体所构成的集合。

总体是指“统计总体”,是一个统计问题研究对象的全体,是具有某种(或某些)共同特征的元素集合。总体的定义是根据所要研究的问题而定的,如果研究某一个班级学生的身高,则总体由该班全体学生的身高组成;如果研究该班级学生的体重,则总体由该班全体学生的体重组成。由此可以看出,总体是由要研究的随机变量的所有可能取值构成的。

总体中的每一个研究对象称为个体。组成总体的个体,至少要具有某一共同的属性或特征,这是区分不同统计总体和辨别总体真假的基本依据。有时,总体的确定还需要事先制订某种规则或标准,对总体及其边界作出详细的说明。在具体研究中,我们往往只关心研究对象的某个指标(即变量),因此,总体又可以是指变量或指标取值的全体。这样,对同一个研究对象,如果要调查几个方面的情况,就可以把它们分成几个总体分别进行研究。例如,上文所提到的,对某个班级的学生,既要了解他们的身高情况,又要研究他们的体重情况,就有身高和体重两个总体。

总体范围的大小,与统计研究的目的和任务有关,也与总体存在的规模有关。按总体中所含个体的数目是否有限,可将总体分为有限总体和无限总体。个体极多或无限多的总体称为无限总体,个体有限的总体称为有限总体。我们也可以从抽象意义上来理解无限总体。例如,研究某种药物对某种疾病的治疗效果是有效还是无效,我们将利用一些发病个体进行药效试验,这部分个体可以看成是来自一个假想总体的样本,这个假想总体由所有发病个体对此药物的治疗效果构成,个体数目可能是无限的,但这个总体并不现实存在,因为并未对所有发病个体用药,但从理论上我们可以对所有发病个体用药。

按照总体的构成是否随时间的变化而变化,可将总体分为动态总体和固定总体;按总体是否由现实存在的研究对象所构成,可将总体分为现实总体和假想总体。例如,研究河南省 2012 年出生婴儿的体重,则总体由河南省 2012 年所有出生婴儿的体重构成,这是一个有限总体、固定总体和现实总体。如果在这个总体的定义中取消时间限制,则这个总体就是一个无限总体(因为时间可以无限地延续下去)和动态总体。

统计分析的目的就是要对总体的特征、不同总体间的差异等作出推断,但由于总体往往很大,而且在实际统计分析中,遇到的总体大多是无限总体、动态总体和假想总体,不可能得到总体中全部个体的数据资料。通常的做法是从总体中按一定方法抽取部分具有代表性的个体,这部分取自总体的个体所构成的集合即为样本。由此也可以看出,统计分析的基本任务就是通过对样本的分析来推断总体。

应该指出的是,在一次调查中,总体是唯一确定的,而样本却是随机变化的。从一个总体中可以抽取多个样本,这取决于样本容量、抽样方式、推断精确度的要求等诸多因素。

4. 正确理解变量的概念

变量是相同性质的事物间表现差异性的某种特征。总体内部的各个体,除了具有某种或某些共同属性外,在调查项目具体表现上存在的差异称为变异。例如,研究某个班级学生的体重情况,体重是调查项目,而每个学生的体重是不完全一样的,这就是变异现象。从某种意义上说,变异是统计存在的前提,若是群体中的各个个体无论哪方面的表现都一样,显然就没有进行统计工作的必要了。统计工作的重要任务,就是通过对存在变异的大量现象的调查研究,剔除偶然性因素的影响,最终得到对总体必然性的认识。因此,变异是相对于调查结果而言的,而变量则是相对于调查项目来说的。广义上讲,凡是能够取不同的值或取值多于一个以上的量均可以称为变量,如身高、体重、温度、酶的活性、叶片叶绿素的含量等都是变量。

5. 正确理解定量资料与定性资料

可以以一定数量形式表达和处理的资料称为定量资料,如小麦籽粒的千粒重、分光光度计的光吸收值读数等。定量资料可以分为连续型变量资料和离散型变量资料。连续型变量资料是指由度量、称量或测量方法得到的数据资料。其观察值不一定是整数,在任何两个观察值之间可以有微小差异的其他数值存在,如籽粒重量、叶片叶绿素含量、人体红细胞含量等。离散型变量资料是指由计数方法获得的数据资料。其仅可以表现为整数形式,如每个培养皿内籽粒发芽的粒数、一定面积内植株发生变异的株数等。定性资料是指

可对其用语言描述和分类的非数量化性状资料。为便于处理或交流,定性资料可按一定的人为原则进行数量化。

6. 正确理解准确性和精确性的概念与区别

准确性是指在调查或试验中某一试验指标或性状的观测值与真值的接近程度;精确性是指调查或试验中同一试验指标或性状的重复观测值彼此的接近程度。准确性不等于精确性。准确性反映了观测值或估计值与真值的接近程度,而精确性则反映了多次重复测定值的变异程度。通常样本所属总体的真值是未知的,只能由样本的统计数来估计和代替,因此准确性无法衡量;而精确性则可以通过样本的统计数据进行衡量。此外,精确性与数据的有效位数有关。有的重复观测值在有效位数较少时(如仅取整数)精确性很高,但如果增加有效位数后,差别就显示出来了。例如,对一个植株个体高度的两次测量值分别为85.8cm和86.3cm,二者之间的差为0.5cm,但如果只取整数,二者就没有差别了。有效位数的多少除人为的取舍外,还与测量仪器的精度有关。例如,一个普通天平,只能称得以克(g)为单位的重量,克以下则不敏感,而一个电子天平可以称0.001g或0.0001g的重量。如果要取得高的精确性,需要测量仪器精度较高,同时还要求保留一定的有效位数,这无疑在一定程度上加大了工作量。

7. 试验中两类误差对试验结果会产生何种影响

试验中的系统误差影响试验结果的准确度,随机误差影响试验结果的精确度。系统误差表现为非处理因素对试验处理效应有规律、有方向的影响,如测量所用的度量衡仪器设备的不标准、田间试验中土壤肥力随地域空间尺度的变化而表现出的梯度变化等。这类误差经过认真研究分析,是可以将其影响消除的。例如,我们可以通过对测量仪器进行校准,对土壤肥力的变化进行局部控制等来进行消除。系统误差对试验结果的影响是一致的或者有一定的规律性,其影响并不会因为处理的改变而改变,一般也不会造成在同一性状重复观测值间精确度的降低。随机误差是各种无规律、相互独立的偶然影响对试验结果综合作用的结果。对随机误差可通过认真细致地试验管理以及适当地试验设计来有效降低,但无法完全消除。随机误差过大的试验一定会使同一性状各重复观测值间的差异变大,即精确度变低。

习题解答

习题 1.1 什么是生物统计学?生物统计学的主要内容和作用是什么?

答 生物统计学是用数理统计的原理和方法来分析及解释生物界各种现象和试验调查资料,是研究生命过程中以样本来推断总体的一门学科。

生物统计学主要包括试验设计和统计分析两大部分内容。其基本作用表现在以下4个方面:①提供整理和描述数据资料的科学方法,确定某些性状和特性的数量特征;②判断试验结果的可靠性;③提供由样本推断总体的方法;④提供试验设计的一些重要原则。

习题 1.2 解释以下概念:总体、个体、样本、样本容量、变量、参数、统计数、因素、水平、处理、重复、效应、互作、试验误差。

答 总体是具有相同性质的个体所组成的集合,是指研究对象的全体。

个体是组成总体的基本单元。

样本是从总体中抽出的若干个个体所构成的集合。

样本容量是指样本个体的数目。

变量是相同性质的事物间表现差异性的某种特征。

参数是描述总体特征的数量。

统计数是描述样本特征的数量。

因素是指试验中所研究的影响试验指标的原因或原因组合。

水平是指每个试验因素的不同状态(处理的某种特定状态或数量上的差别)。

处理是指对受试对象给予的某种外部干预(或措施)。

重复是指在试验中,将一个处理实施在两个或两个以上的试验单位上。

效应是由处理因素作用于受试对象而引起试验差异的作用。

互作是指两个或两个以上处理因素间的相互作用产生的效应。

试验误差是指试验中不可控因素所引起的观测值偏离真值的差异,可以分为随机误差和系统误差。

习题 1.3 随机误差与系统误差有何区别?

答 随机误差也称为抽样误差或偶然误差,它是由于试验中许多无法控制的偶然因素所造成的试验结果与真实结果之间的差异,是不可避免的。随机误差可以通过试验设计和精心管理设法减小,但不能完全消除。

系统误差也称为片面误差,是由于试验处理以外的其他条件明显不一致所产生的带有倾向性的或定向性的偏差。系统误差主要由一些相对固定的因素引起,在某种程度上是可控制的,在试验过程中是可以避免的。

习题 1.4 准确性与精确性有何区别?

答 准确性也称为准确度,是指在调查或试验中某一试验指标或性状的观测值与真值接近的程度。精确性也称为精确度,是指调查或试验中同一试验指标或性状的重复观测值彼此的接近程度的大小。

准确性是说明测定值对真值符合程度的大小,用统计数接近参数真值的程度来衡量。精确性是反映多次测定值的变异程度,用样本中的各个变量间的变异程度的大小来衡量。

自我测验

一、填空

1. 变量按其性质可以分为_____变量和_____变量。
2. 样本统计数是总体_____的估计值。
3. 生物统计学是研究生命过程中以样本来推断_____的一门学科。

- 4. 生物统计学的基本内容包括_____、_____两大部分。
- 5. 统计学的发展过程经历了_____、_____、_____ 3 个阶段。
- 6. 生物学研究中,一般将样本容量_____称为大样本。
- 7. 试验误差可以分为_____、_____两类,通过矫正或实施控制可予以消除的一类误差是_____,而通过合理的试验设计和精心管理可以减小,但无法完全消除的一类误差是_____。

二、判断

- () 1. 对于有限总体不必用统计推断方法。
- () 2. 资料的精确性高,其准确性也一定高。
- () 3. 在试验设计中,随机误差只能减小,而不可能完全消除。
- () 4. 统计学上的试验误差,通常是指随机误差。

三、名词解释

连续变量 非连续变量 定量变量 定性变量 因素 试验单位

第二章 试验资料整理与特征数计算

目的要求

- (1) 掌握抽样调查的概念和方法。
- (2) 熟悉不同类型资料的整理和相关统计图表的制法。
- (3) 掌握几种常用的平均数和变异数的基本概念及计算方法。

内容提要

试验资料的搜集与整理是数据资料处理的首要环节。根据生物的性状特征,试验资料可分为数量性状资料和质量性状资料两类。数量性状资料是由计数或测量的方法得到的,又分为计数资料(非连续变量资料)和计量资料(连续变量资料)。质量性状资料(属性资料)经常需要经过数量化再进行统计分析。

搜集试验资料的方法有调查和试验,而调查又可分为普查和抽样调查两种方式。其中,抽样调查是常用的调查方法。它是从全部调查研究对象中,抽选一部分单位进行调查,并对全部调查研究对象作出估计和推断的一种非全面调查方法。根据研究情况的不同,常用抽样调查方法有随机抽样、顺序抽样和典型抽样等。

调查或试验取得数据资料后,需对其进行检查、核对,确保资料正确无误后即可进行资料整理。在资料整理分组时,计数资料常用单项式分组法,计量资料常用组距式分组法。根据分组的资料,制作次数(频率)分布表和次数(频率)分布图。常用的统计图形有适合于计数资料和属性资料的条形图、饼图,适合于计量资料的直方图、多边形图,反映变量间相关性及变化趋势的散点图等。通过制作统计图表可以定性地反映资料的特征,但要定量描述其特征,还要进一步计算资料的特征数。

试验资料具有集中性和离散性两个基本特征。平均数是反映集中性的特征数,变异数是反映离散性的特征数。常用平均数包括算术平均数、中位数、众数和几何平均数等,算术平均数具有离均差之和等于零及离均差平方和为最小等基本性质,可以用直接算法、减去(加上)常数法和加权法来计算。

常用变异数包括极差、方差、标准差和变异系数等。极差是资料中最大值和最小值之差,计算简单,但只能反映数据的最大波动范围;方差等于观测值离均差的平方和除以其自由度,可以反映出资料中每一个观测值的变异;标准差是方差的平方根,是表示资料平均变异程度的一项重要指标;用标准差除以其平均数即为变异系数,变异系数是变量的相对变异量,可以进行平均数相差悬殊或单位不同的资料间变异程度的比较。

难点评析

1. 正确理解随机化的意义

随机化是指被研究的样本从总体中任意抽取,且总体中每一个观察单位都有同等机会被抽取为样本个体的可能性。在试验研究中,要求各比较组之间除了处理因素外,其他非处理因素最好完全一致或均衡,事实上我们很难达到这一点,只有采用随机化分组法,才可能使各组的非处理因素尽量趋于一致或均衡。随机化的意义表现在:① 避免有意或无意夸大或缩小组间差别而使试验结果产生偏差;② 抽样研究的理论和统计分析方法是以随机化为基础的,要求试验设计遵循随机原则。

2. 理解抽样调查的作用

抽样调查是一种非全面调查,它是从总体中抽取样本,以样本的特征对总体进行推断的一种调查方法。可以从以下几个方面理解抽样调查的作用。

(1) 无限总体的调查。总体的单位数无限多,由于不可能把它们一一列出,采用全面调查显然是不可行的,只能进行抽样调查。例如,空气质量的调查和作物病害的调查等就是如此。有些调查对象,尽管包含的单位数有限,但数目巨大,难以列出清单,实际工作往往将它们作为无限总体来处理,采用抽样调查进行研究。

(2) 一些破坏性调查或试验,如植株或动物个体的解剖试验等,采用全面调查是不可行的。

(3) 一些大规模的调查,如环境资源的调查、疾病发生情况的调查等,从理论上讲是可以进行全面调查的,但实行起来却较困难,甚至不可能,需要通过抽样调查进行。

(4) 抽样调查的费用较低。同普查相比,抽样调查仅从占总体很少一部分的单位中搜集数据,显然其费用与普查相比要低得多。

(5) 抽样调查的速度较快。搜集和综合样本资料与搜集和综合全面调查资料相比,工作量小,因此在时间上通常要快很多。在对资料的时效性要求比较高时,抽样调查即表现出明显的优势。

(6) 应用范围广泛,灵活性强。随着抽样理论的完善,抽样调查的应用范围日趋广泛,既适应于连续型现象,也适应于不连续型现象;既适用于经常性调查,也可用于不定期的一次性调查(根据需要随时可以进行调查)。另外,抽样调查既可用于大范围的调查,也可用于小范围的调查;既可以由组织、单位举行,也可以由个人或几个人独立完成。

(7) 抽样调查的准确度较高。抽样调查的工作量较小,因此能够组织起素质较高的调查队伍,容易进行严格仔细的检查监督,保证资料处理的及时性和完整性,从而也就能获得更加准确的调查结果。

3. 抽样调查的特点

抽样调查有以下 3 个特点。

(1) 总体的各单位都有一个指定的概率被抽取。其中,每个单位有同等概率被抽中的,称为等概率抽样;被抽中的概率不相等的,称为不等概率抽样。

(2) 随机原则。不论是等概率抽样还是不等概率抽样,在抽取样本单位时必须通过一个或几个随机的步骤进行,而不得采取主观、有意的选择方式。

(3) 推断总体。可用样本的估计值来推断总体的参数,并可以用一定的概率来保证将误差控制在规定的范围内。

4. 正确理解抽样误差的作用及其影响因素

抽样误差在抽样调查中具有重要的意义,通过计算抽样误差,可以反映估计量的好坏,同时也可以正确认识估计结果的精确程度。抽样误差可以计算,也能够加以控制。在抽样过程中以下因素影响抽样误差的大小。① 总体的差异程度:总体的差异程度越大,抽样误差也随之增大;反之,抽样误差越小。② 样本容量:抽取的样本单位数目多,抽样误差小;抽取的样本单位数目少,抽样误差就会增大。③ 样本单位的抽取方式:重置抽样的误差一般大于非重置抽样的抽样误差。④ 抽样调查的组织方式:不同的调查组织方式,如简单随机抽样、分层抽样、整群抽样等,所产生的抽样误差一般是不同的。⑤ 估计量的选择:不适当的估计量会产生较大的抽样误差。

5. 如何对抽样方法进行评价

通常从两个方面来评价抽样方法,其一是抽样估计精确度的大小;其二是调查费用的多少。抽样调查费用的多少与样本容量、样本单位的分布范围、抽样调查的组织方式、资料搜集的难易程度、调查项目的多少均有密切的关系。一个好的抽样方法应该是抽样估计精确度高而调查费用较少的抽样方法。

6. 如何确定样本容量

简单随机抽样时,确定样本容量需要考虑以下因素。① 抽样估计的精度要求:精度要求高的应多抽一些样本,精度要求低的可以少抽一些样本。② 总体的差异程度:总体差异程度大,为了必要的抽样效果,可考虑多抽一些样本,否则可以少抽一些样本。③ 抽样方式:由于重置抽样的误差比非重置抽样的大,因此,重置抽样时的样本容量应比非重置抽样的要稍大一些。④ 推断的概率把握程度:提高推断的概率把握程度,必须增加样本数,如果对抽样估计的置信水平要求不高,此时可使样本容量少一些。⑤ 调查费用的多少:如果调查费用较充足,可适当增加样本单位;如果调查费用相对紧张,则应减少样本单位。应该注意的是,虽然有时这些影响因素难以同时得到满足,但是应尽可能兼顾。

7. 为什么要研究频数资料

对于样本资料,特别是大样本的数据资料,通过研究其频数(率)分布,可以从杂乱无章的数据中,快速、简单、直观地获得资料的主要特征,如通过极差(全距)可以了解数据的变异程度,也可以了解资料的离散性(即数据资料远离其中心的趋势);通过频数分布表或

分布图可以了解数据资料的集中性,获得大多数资料的分布区间,即了解资料的集中性;通过频数的分布情况也可以了解资料的变化趋势,即随着变量的变化频数是如何变化的,资料的分布是对称的还是偏态的。

8. 关于自由度

在方差计算公式中,样本的方差不以样本容量 n 作为除数,而是以自由度 $n-1$ 作为除数。这是因为我们通常所掌握的资料,大多不知其总体平均数 μ 的数值,常用样本平均数 \bar{x} 来估计 μ 。由于 \bar{x} 与 μ 总有差异,根据平均数的基本性质——离均差平方和为最小,则有 $\sum(x-\bar{x})^2 < \sum(x-\mu)^2$ 。因此由公式 $\frac{\sum(x-\bar{x})^2}{n}$ 计算出来的方差对总体所作的估计是偏小的。用 $n-1$ 来代替 n ,去除离均差平方和,可以避免偏小估计的问题,从而实现样本方差对总体方差的无偏估计。

在统计上,自由度是指样本内独立而能自由变动的观测值的个数。在计算 n 个观测值的样本方差时,每个 x 与 \bar{x} 比较,虽有 n 个离均差,但只有 $n-1$ 个是自由变动的,最后一个离均差由于受 $\sum(x-\bar{x})=0$ 的限制不能自由变动。例如,5个观测值的样本,如果4个离均差为2、3、1、-2,则第5个离均差一定为-4;如果4个离均差为1、-3、4、-4,则第5个离均差一定为2,这样才能使 $\sum(x-\bar{x})=0$ 成立。由于能自由变动的离均差是4个,故自由度为4,即自由度为 $n-1$ 。在计算其他统计数时,如果受到 k 个条件的限制,则其自由度应为 $n-k$ 。

9. 关于标准差

在标准差的特性中,“如果将各观测值加上或减去一个常数 a ,其标准差不变;如果将各观测值乘以或除以一个常数 a ,则所得到的标准差扩大或缩小了 a 倍”可以通过以下公式推导得到。

已知一组观测值,其标准差 $s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$ 。若将每一个观测值 x_i 加上常数 a ,把 $x_i + a$ 代入标准差公式中,则有

$$\begin{aligned} s' &= \sqrt{\frac{\sum (x+a)^2 - \frac{[\sum (x+a)]^2}{n}}{n-1}} \\ &= \sqrt{\frac{\sum x^2 + 2\sum ax + \sum a^2 - \frac{(\sum x + \sum a)^2}{n}}{n-1}} \\ &= \sqrt{\frac{\sum x^2 + 2a\sum x + na^2 - \frac{(\sum x)^2 + 2na\sum x + n^2a^2}{n}}{n-1}} \end{aligned}$$