社会统计学
(Social Statistics)

科学只有当它利用了数学的时候，它才达到了完善的程度。

——马克思

对于追求效率的公民而言，统计思维总有一天会和读写能力一样必要。

——H.G.Wells

教材及参考书目

社会统计学，张彦，高等教育出版社，2005

社会统计学，张彦，南京大学出版社，1997

社会统计学（第八版），布莱洛克，社会科学文献出版社

社会统计学（重排本），卢淑华，北京大学出版社，2002

社会研究的统计分析，李沛良，社会科学文献出版社

17 世纪以前，社会统计主要局限于对事物进行原始的调查登记和简单的计算汇总。

如大禹时的九州表，明初的黄册和鱼鳞册；古埃及、古希腊、古罗马在公元前 400 年就建立的出生、死亡登记制度。

17 世纪后，产生了以工业、农业、贸易、交通等方面统计为主的社会经济统计。

国势学派

政治算术学派

数理统计学派

1.国势学派

代表人物是康令（1606~1681）和阿亨瓦尔（1719~1772）。1749 年，阿亨瓦尔根据拉丁文“Status”意大利文 Stato 和 Statista 及德文 Statis 等字根创造出 “Statistik”这个新词，原意指“国家显著事项的比较和记述”。

国势学派可谓“有名无实”的学派：只用文字记述，不用数字计量。它又称记述学派和历史学派。

2. 政治算术学派

格朗特 1662 年在其《自然和社会观察》一书中，从宗教管理、商业、气候、疾病等方面，对当时伦敦人口的出生率、死亡率和性比例等方面进行了综合的统计分析。

威廉 •配第 1667 年在其《政治算术》一书中，运用有关人口、土地税收和国家收入等方面的数字资料，对英国、荷兰的经济实力进行比较，首创了一种数字对比分析的方法。“即用数字、重量、尺度来表达自己想说的问题。”

与国势学派相对应，政治算术学派可谓“有实无名”的学派

3.数理统计学派

凯特勒（1796~1896） 首先将概率论原理引入到社会现象的研究，在《社会物理学》,《道德统计》、《论人类》等书中，他认识到人类的社会活动服从于一定规律，并发现这种规律只有通过大量观察才能被人们所认识。

凯特勒被称为现代统计学之父。1867 年，一门兼有数学和统计学双重性质的学科被命名为“数理统计学”。

1886 年，高尔顿：相关指数

1900 年，皮尔逊：卡方检验，复相关计算

1928 年，戈塞特 t；费舍 F

1950 S，拉扎斯菲尔德：自动化处理

1966 年，斯坦福：SPSS

4.社会统计学派

凯特勒的另一个重要贡献，是他把政治经济学、数学和当时政府统计工作的方法结合在一起，建立了一个专门研究社会现象的统计学派。后来这个学派传到德国，就出现了以克尼斯(1821—1898)、梅尔(1841—1 923)和恩格尔

(1821—1896)为代表的德国社会统计学派。

第一次世界大战前后，随着社会统计学派的中心逐步向英、美等国转移，社会统计学与社会学的关系日益明确。

1900 年，马约 •史密斯《统计学和社会学》。

1920 年，史特威 •恰平《实地调查与社会研究》。恰平还著有《社会学中的科学方法》等书。

二次大战后，社会统计学在广义和狭义两方面的实践意义逐步得到了人们的公认。

20 世纪 60 年代以来，西方发达资本主义国家先后都制定了社会发展计划。 20 世纪 60 年代首先在美国掀起了一个颇有声势的 "社会指标运动"。

1966 年，雷蒙 •布埃尔提出用社会指标的方法解决社会分析和社会规划的基本理论，出版了《社会指标》一书。

1976 年，经互会《社会统计基本指标体系》

1976 年，OECD 《社会生活质量的计量》

1982 年，国家统计局成立社会统计司

1983 年,《中国社会统计资料》首次公开出版

标题部分

1、标题置于表格正上方

2、总标题所示要点与表中项目一致，在需要时还应表明资料所属的时间和地区

3、表次：左 ； 单位：右

4、对分页的同一表格，在每页上端都要写标题，加（续一）、（续二）

栏目部分

1、先局部后整体

2、若栏目较多，可加以编号；统计数字间有计算关系的，可用计算式表达。

线格部分

1、统计表上下两端线应以粗线或双细线标划，表的左右两侧开口。

2、各栏间用直线标划，大项目间线条较粗，小栏目线条较细；各行间不必划线条。

数字部分

1、表中数字要对准位数。

2、不存在某数字时，用"——表示；缺少某项数字时，用"……表示。

3、数字较大时，加分位点。

其他规则

1、资料来源写在表格下方。

2、有说明解释需要时，在表下方注释。

3、单位有数种时而不能在表右上角划一标注时，分两种情况处理。

1.单项式变量数列——数列中每一组的变量值只有一个。单项数列用于离散变量整数值变动幅度较小时。

某社区各户人口数统计表

对于等距分组且为闭口组的情况，确定组距已有某些数学公式可供参考，但最佳决定还是依据常识和数列使用的目的而定。一般地说，组距应不小于可以忽略的数值之差。

注意，在资料被整理成数列时，全距可适当放大(但不能缩小)，以便组数(或组距)取整数值。

异距分组

异距分组主要在变量变动并不是均匀的、有急剧上升或突然下降之类情况发生时考虑。

标准组距频数的换算方法：

（1）选定某一合适的组距为标准组距；

（2）用标准组距除以各组组距，得到折合系数；

（3）将各组的折合系数乘以各组的频数。

累计频数（F）

向上累计——以变量数列首组的频数为始点，逐个累计各组的频数，展示小于该组上限的频数和。

向下累计——以变量数列末组的频数为始点，逐个累计各组的频数，展示大于该组下限的频数和。

频数分布不但可以用统计表的形式表现，也可以用统计图的形式表现。用统计图表示频数分布，较之用统计表，要直观便捷得多。但缺点是不及统计表精确。统计图的种类很多，本节仅就与频数分布数列相衔接的统计图加以介绍。

根据编制好的频数分布数列，可以绘制出相应的统计图，最常用的有频数分布直方图、折线图、曲线图以及累计频数分布曲线。

具体方法是：先画直角坐标系，横轴代表分组或各组组限，纵轴代表各组频数或频率，然后再根据相应的分配数列作图。

条件下，很显然各矩形的面积与其高度成正比。因此，各矩形的面积同样可以用来表示各组的频数或频率，而且看起来更形象直观。如果取各矩形的总面积为 1，各矩形的面积必定等于各组的相对频数。

直方图(Histograms)

直方图是用矩形(或长条)的高度来表示数列各组的频数或频率。对于定类变量和定序变量的分组，矩形(或长条)的宽度是没有意义的，各矩形之间要留出一定的空隙；对于定距变量(和定比变量)的分组，矩形的宽度表示各组组距，各矩形之间一般不留空隙。在等距分组的其实，在频数分布图中，用面积来理解频数分布状况更合适。

比如直方图，当处理异距分组时应该用矩形面积而不是用矩形高度来显示频数分布。

下面是根据表 3．15 绘制出的两个直方图。左图用矩形高度来表示各组频数就会产生错觉。右图是按照标准组距频数作出来的，用矩形面积来表示各组频数就避免了不必要的错觉。以后当我们接触正态曲线等曲线后，将进一步体会到用面积来表示频数分布的好处。

折线图（Polygon）

表示频数分布的另一种图形是频数多边形图，简称折线图。直接把直方图各矩形顶部的中点用直线连接起来，并把原来的矩形抹掉，就得到频数多边形图。

当变量数列中的组数愈加增多，变量值也非常多时，折线图会逐步过渡到平滑曲线。频数分布曲线图实质上是对应于连续变量的频数分布的函数关系图。

下表是诺贝尔获奖者的年龄分布表。(1)请根据数据制作直方图和折线图；（2）将折线图修匀为一条曲线图，并描述该曲线的特点。

常见曲线图类型

●逻辑斯蒂曲线：变量值分布的次数随变量值增大 而增多或相反，但有上限。

●累计频数分布曲线

显然，累计频数分布曲线只有两种形状：或持续增长的或持续减少的。这分别取决于向上累计或向下累计。累计频数分布曲线一般都呈逻辑斯谛曲线形，其斜率最大的地方对应于频数最大的组，其水平的地方对应于空组。

基尼系数的计算公式，可以根据定义，用求三角形和梯形面积的方法，很简单地推导出来，即第四章　集中趋势测量法。下面是一个小故事：

一个人到某公司求职，经过调查，得出关于该公司工资的一些数据，如果是你，应该如何选择？

我们有三种方法选择集中趋势：

（1）根据频数：哪个变量值出现次数越多，就选择哪个变量值，比如民主决策的表决机制。

（2）根据居中：比如一个城镇居民的生活水平，居中的是小康家庭，那么就用小康家庭来代表该城镇的生活水平。

（3）根据平均：用平均数来代表变量的平均水平。

关于集中趋势的一个故事

吉斯莫先生有一个小工厂，生产超级小玩意儿。 管理人员由吉斯莫先生、他的弟弟、六个亲戚组成。工作人员由 5 个领工和 10 个工人组成。工厂经营得很顺利，现在需要一个新工人。现在吉斯莫先生正在接见萨姆，谈工作问题。

吉斯莫：我们这里报酬不错。平均薪金是每周 300 美元。你在学徒期间每周得 75 美元，不过很快就可以加工资。萨姆工作了几天之后，要求见厂长。

萨姆；你欺骗我！我已经找其他工人核对过了，没有一个人的工资超过每周 100 元。平均工资怎么可能是一周 300 元呢？

吉斯莫：啊，萨姆，不要激动。平均工资是 300 元。我要向你证明这一点。

吉斯莫：这是我每周付出的酬金。我得 2400 元，我弟弟得 1000 元，我的六个亲戚每人得 250 元，五个领工每人得 200 元，10 个工人每人 100 元。总共是每周 6900 元，付给 23 个人，对吧？

萨姆：对，对，对！你是对的，平均工资是每周 300 元。可你还是蒙骗了我。

吉斯莫；我不同意！你实在是不明白。我已经把工资列了个表，并告诉了你，工资的中位数是 200 元，可这不是平均工资，而是中等工资。

萨姆：每周 100 元又是怎么回事呢？

吉斯莫：那称为众数，是大多数人挣的工资。

吉斯莫：老弟，你的问题是出在你不懂平均数、中位数和众数之间的区别。

萨姆：好，现在我可懂了。我……我辞职！

第一节 算术平均数(MEAN)

注意：对求和符号，此时流动脚标的变动范围是 1, 2, 3, …,ℕ 是总体单位数。

[例] 求 74、85、69、9l、87、74、69 这些数字的算术平均数。

注意：对求和符号，此时流动脚标的变动范围是 1,2,3 …,m 是组数，而不是总体单位数。

很显然，算术平均数不仅受各变量值(X)大小的影响，而且受各组单位数(频数)的影响。由于对于总体的影响要由频数(f戈小所决定，所以 f 也被称为权数。值得注意的是，在统计计算中，权数不仅用来衡量总体中各标志值在总体中作用，同时反映了指标的结构，所以它有两种表现形式：绝对数（频数）和相对数（频率）。这样一来，在统计学中，凡对应于分组资料的计算式，都被称为加权式。

[例] 求下表(单项数列)所示数据的算术平均数 。

对于组距数列，要用每一组的组中值权充该组统一的变量值。

[例] 求下表所示数据的的算术平均数

第二节 中位数（Median)

例 求 54，65，78，66，43 这些数字的中位数。

例、求 54，65，78，66，43，38 这些数字的中位数。

（2）组距数列按中位数所在组的下限：按中位数所在组的上限：

4. 四分位数

中位数所有单位被等分为两部分，因而被称为二分位数。类似于求中位数，我们还可求出四分位数、十分位数、百分位数。

将总体中的各单位分割成相等的四部分，则这三个分割的变量值就是四分位数。若以 Q1 、Q2 、Q3 分别代表第一、第二、第三四分位数。Q2 即中位数，Q1 、Q3 的算法分别是

请从下表中指出第一四分位数和第三四分位数求出下表中的第一四分位数和第三四分数

第三节 众数(Mode)

1. 对于未分组资料

直接观察

首先，将所有数据顺序排列；然后，只要观察到某些变量值(与相邻变量值相比较)出现的次数(或频数)呈现 “峰”值，这些变量值就是众数。

2. 对于分组资料

单项式：观察频数分布 (或频率分布)

组距式：

求下表中的众数

(1) 众数仅受上下相邻两组频数大小的影响，不受极端值影响，对开口组仍可计算众数；

(2) 受抽样变动影响大；

(3) 众数不唯一确定。

(4) 众数标示为其峰值所对应的变量值，能很容易区分出单峰、多峰。因而具有明显偏态集中趋势的频数分布，用众数最合适。

第四节 几何平均数、调和平均数（了解）

1. 几何平均数 Mg (geometric mean)

N 个变量值连乘积的 N 次方根。(不能有变量值为 0)。适用于：(1) 计算某种比率的平均数；(2) 计算大致具有几何级数关系的一组数字的平均数，如经济指标的平均发展速度。

应该指出，用以计算几何平均数的各项数值必须大于 0，否则就不能计算几何平均数或计算结果无实际意义。

[例] 求 3，9，27，81，243 这些数字的几何平均数。

2. 调和平均数 Mh (harmonic mean)

N 个变量值倒数算术平均数的倒数，也称倒数平均数。适用于：掌握的情况是总体标志总量而缺少总体单位数的资料时。

简单调和平均数

加权调和平均数

3. 各种平均数的关系

(1) 当总体呈正态分布时：

(2)当总体呈偏态分布时：中位数总在均数和众数之间

正偏： 负偏：

（注： 和 合称位置平均数）

(3) 皮尔逊发现，在钟形分布的偏态不大显著时， 、 、 三者大致构成一个比较固定的关系：

第五章 离中趋势测量法

例如有 A 、B 、C 、D 四组学生各 5 人的成绩如下：

A 组：60 ，60，60，60，60

B 组：58，59，60，61，62

C 组：40，50，60，70，80

D 组：80，80，80，80，80

数据显示，平均数相同，离势可能不同；平均数不同，离势可能相同。

变异指标用以反映总体各单位标志值的变动范围或参差程度，与平均指标相对应，从另一个侧面反映了总体的特征。

变异指标如按数量关系来分有以下两类；

凡用绝对数来表达的变异指标，统称绝对离势;

凡用相对数来表达的变异指标，统称相对离势;

第一节　全距与四分位差

1.全距(Range)

[例]　求 74，84，69，91，87，74，69 这些数字的全距。

[解]　把数字按顺序重新排列：69，69，74，74，84，87，91，显然有

R =Xmax －Xmin ＝91—69＝22

2.四分位差(Quartiledeviation)

第三四分位数和第一四分位数的半距。避免全距受极端值影响大的缺点。

第二节　平均差(Mean absolute deviation)

要测定变量值的离中趋势，尤其是要测定各变量值相对于平均数的差异情况，一个很自然的想法就是计算各变量值与算术平均数的离差。平均差是离差绝对值的算术平均数。(mean deviation)

1.对于未分组资料

A •D=

2.对于分组资料　　A •D=

3.平均差的性质

[例 1]　试分别以算术平均数为基准，求 85，69，69，74，87，91，74 这些数字的平均差。

[例 2] 试以算术平均数为基准，求下表所示数据的平均差。

第三节　标准差（standard deviation)

求 72、81、86、69、57 这些数字的标准差。

2.　对于分组资料

[例]　调查大一男生 60 人的身高情况如下表所示，求他们身高的标准差。

[解]　因为是分组资料，计算标准差运用加权式，并

参见下表

标准差是反映总体各单位标志值的离散状况和差异程度的最佳测度。

（1）以算术平均数为基准计算的标准差比以其他任何数值为基准计算的标准差要小。 “最小二乘方”性质——各变量值对算术平均数的离差的平方和，必定小于他们对任何其他数偏差的平方和。

（2）它将总体中各单位标志值的差异全包括在内，受抽样变动影响小。但在受极端值影响以及处理不确定组距方面，缺点同算术平均数。

值得注意的是，在推论统计中我们将发现，方差是比标准差更有理论价值的概念。所谓方差，即标准差的平方，它直接写成。也常被称为变异数。

4.标准分(standard score)

以离差和标准差的比值来测定变量　　与　　的相对位置。使原来不能直接比较的离差标准化，可以相互比较，加、减、平均。

Z 分数也有标准正态变量之称。按 Z 值大小编制出的正态分布表，其用途十分广泛。

Z 分数的性质：

第四节　相对离势

上述各种反映离中趋势的变异指标，都具有和原资料相同的计算单位，称绝对离势。但欲比较具有不同单位的资料的参差程度，或比较单位虽相同而均值不相同的资料的参差程度，离势的绝对指标则很可能导致某些错误结论。所以，我们还得了解和学习相对离势。

全距系数

全距系数是众数据的全距与其算术平均数之比，其计算公式是

平均差系数

平均差系数是众数据的平均差与其算术平均数之比，其计算公式是

标准差系数

标准差系数是众数据的标准差与其算术平均数之比，其计算公式是

异众比率能表明众数所不能代表的那一部分变量值在总体中的比重。

2. 异众比率

所谓异众比率，是指非众数的频数与总体单位数的比值，用 V·R 来表示

其中： 为众数的频数； 是总体单位数

例 1：某项调查发现，现今三口之家的家庭最多（32% ），求异众比率。某开发商根据这一报导，将房屋的户型大部分都设计为适合三口之家居住的样式和面积，你认为如何呢？

例 2：设为测体重，得到成人组和婴儿组各 100 人的两个抽样总体。成人组平均体重为 65 千克，全距为 10 千克；婴儿组平均体重为 4 千克，全距为 2.5千克。能否认为成人组体重的离势比婴儿组体重的离势大？

例 3：对一个群体测量身高和体重，平均身高为 170.2厘米，身高标准差为 5.30厘米；平均体重为 70 千克，体重标准差为 4.77千克。比较身高和体重的离散程度。

3. 偏态系数

偏斜系数是以标准差为单位的算术平均数与众数的离差，其取值一般在 0 与土 3 间。

偏斜系数为 0 表示对称分布，偏斜系数为或则表示极右或极左偏态。

第六章 概率与概率分布

第一节 基础概率

概率论起源于 17 世纪，当时在人口统计、人寿保险等工作中，要整理和研究大量的随机数据资料，这就需要一种专门研究大量随机现象的规律性的数学。

参赌者就想：如果同时掷两颗骰子 ，则点数之和为 9 和点数之和为 10 ，哪种情况出现的可能性较大？

例如 17 世纪中叶，贵族德¡ㄖ梅尔发现：将一枚骰子连掷四次，出现一个 6 点的机会比较多，而同时将两枚掷 24 次，出现一次双 6 的机会却很少。

概率论的创始人是法国的帕斯卡(1623—1662)和费尔马(1601—1665)，他们在以通信的方式讨论赌博的机率问题时，发表了《骰子赌博理论》一书。棣莫弗(1667—1754)发现了正态方程式。同一时期瑞士的伯努利(1654一1705)提出了二项分布理论。1814 年，法国的拉普拉斯(1749—1827)发表了《概率分析论》，该书奠定了古典概率理论的基础，并将概率理论应用于自然和社会的研究。此后，法国的泊松(1781—1840)提出了泊松分布，德国的高斯(1777—1855)提出了最小平方法。

在统计学中，我们把类似掷一枚硬币的行为（或对某一随机现象进行观察）称之为随机试验。随机试验必须符合以下三个条件：①它可以在相同条件下重复进行；②试验的所有结果事先已知；③每次试验只出现这些可能结果中的一个，但不能预先断定出现哪个结果。

[例] 对掷一颗骰子的试验，我们研究如下事件：①A 为 “点数是 3”；②B 为 “出现奇数点”；③C 为 “出现点数不超过 6”；④D 为 “点数是 7”。

[解] 因为 Ω＝{1，2，3，4，5，6}，所以

①A ＝{3} ，为简单事件；

②B ＝{1，3，5}，为复合事件；

③C ＝{1，2，3，4，5，6}，为必然事件；

④D ＝{7}，为不可能事件。

2. 事件之间的关系

（1）事件和（Or conjunction)——事件 A 与事件 B 至少有一个事件发生所构成的事件 C 称为 A 与 B 的事件和，记作

（2）事件积(As-well-as conjunction)事件 A 与事件 B 同时发生所构成的事件 C 称为 A 与 B 的事件积，记作

（3）事件的包含与相等——事件 A 发生必然致事件 B 发生，则称为 B 包含 A 记作

（4）互斥事件——事件 A 和事件 B 不能同时发生，则称 B 和 A 是互斥事件，或互不相容事件，记作

（5）对立事件——事件 A 与事件 B 是互斥事件，且在一次试验中必有其一发生，称 A 与 B 为对立事件（逆事件），记作

（6）相互独立事件——事件 A 的发生与事件 B 是否发生毫无关系，称 A 与 B 为相互独立事件，记作

之间的两关系随机事件

[例] 掷两枚均匀的硬币， ①求“两枚都朝上”的概率； ②求“一枚朝上，一枚朝下”的概率。

这样对于含有 m 个样本点的事件 A，其出现的概率为

4. 经验概率

求算概率的另一途径是运用频率法。设想有一个与某试验相联系的事件 A，把这个试验一次又一次地做下去，每次都记录事件 A 是否发生了。假如做了 n 次试验，而记录到事件 A 发生了 m 次（即成功 m 次），则频数与试验次数的比值，称作次试验中事件 A 发生的频率

显然，频率具有双重性质：随机性和规律性.

当试验或观察次数趋近于无穷时相应频率趋于稳定，这个极限值就是用频率法所定义的概率，即

频率稳定到概率这个事实，给了“机会大小”即概率一个浅显而说得通的解释，这在统计学上具有很重要的意义。坚持这种观点的统计学派也就被称为频率学派。

2.加法规则

如果事件 A 和事件 B 互斥，那么

如果 A 和 B 是任何事件(不一定互斥)，加法规则更普通地表示为如下形式

[例]从一副普通扑克牌中抽一张牌，求抽到一张红桃或者方块的概率。

[例] 在一副 52 张扑克牌中，求单独抽取一次抽到一张红桃或爱司的概率。

加法规则可推广到对两个以上的事件，若事件 A，B，C…K 都互斥，那么有

P (A或 B 或 C…或 K) ＝P(A)+P(B)+P(C)… +P(K)

[例] 根据上海市职业代际流动的统计，向下流动的概率是 0.07，静止不动的概率是 0.6，求向上流动的概率是多少？

[例] 为了研究父代文化程度对子代文化程度的影响，某大学统计出学生中父亲具有大学文化程度的占 30%，母亲具有大学文化程度的占 20%，而双方都具有文化程度的占有 10%，问从学生中任抽一名，父代至少有一名具有大学文化程度的概率是多少？

3.乘法规则

式中符号 和 代表条件概率。 应理解为，“在 B 已经发生条件下 A 发生的概率”。条件概率的意思是，A 发生的概率可能与 B 是否发生有关系。换言之，B 已经发生时 A 发生的概率可能有别于 B 没有发生时 A 发生的概率。

理解统计独立的概念，对于灵活运用概率的乘法规则很重要。现在用条件概率来加以表达，统计独立是指

若 A 和 B 在统计上相互独立(无关)，这时乘法规则可以简化为

[例]假定有下列 3000 个社区的数据，如果随机地从这个总体中抽取一个社区，得到一个中等的而且犯罪率低的社区的概率是多少?

[例]假定数据变动如下，随机地从这个总体中抽取一个社区，得到一个中等的而且犯罪率低的社区的概率又是多少?

[例] 根据统计结果，男婴出生的概率是22/43，女婴出生的概率是21/43，某单位有两名孕妇，问两名孕妇都生男婴的概率是多少？都生女婴的概率是多少？其中一男一女的概率是多少？

[例] 某居民楼共20户，其中核心家庭为2户，问访问两户都是核心家庭的概率是多少？问访问第二户才是核心家庭的概率是多少？

[例] 为了研究父代文化程度对子代文化程度的影响，某大学统计出学生中父亲具有大学文化程度的占30%，母亲具有大学文化程度的占20%，而双方都具有文化程度的占有10%，问从学生中任抽一名，父代至少有一名具有大学文化程度的概率是多少？

在抽样方法中还经常涉及到回置抽样和不回置抽样。如前所述，所谓回置抽样，就是抽取的单位登记后又被放回总体中去，然后再进行下一次抽取。使用回置抽样法，先后两次抽取是彼此独立的。因为每一次抽取后抽取到的单位都得返还，总体保持不变，前一次的结果不可能影响到后一次。所谓不回置抽样，就是不再把抽取到的单位退还总体。这样先后两次抽取就不再独立了，必须使用条件概率的概念。

[例]用回置法从一幅普通扑克牌抽取两次，计算得到两张爱司的概率。

例：用不回置法从一幅普通扑克牌抽取两次，计算得到两张爱司的概率。

在抽样方法中还经常涉及到回置抽样和不回置抽样。如前所述，所谓回置抽样，就是抽取的单位登记后又被放回总体中去，然后再进行下一次抽取。使用回置抽样法，先后两次抽取是彼此独立的。因为每一次抽取后抽取到的单位都得返还，总体保持不变，前一次的结果不可能影响到后一次。所谓不回置抽样，就是不再把抽取到的单位退还总体。这样先后两次抽取就不再独立了，必须使用条件概率的概念。

用不回置法从一幅普通扑克牌抽取两次，计算得到两张爱司的概率。

4. 排列和样本点的计数

要正确解决概率问题，往往光考虑乘法规则还不够，还要同时考虑使用加法规则。一般最简单的做法是：首先确定一种符合要求

的排列方式并计算它们发生的概率，然后再考虑还有没有其他同样符合要求的排列方式。如果存在着其他实现方式，并且都具有相同的概率，就可以简单地把排列方式数与以某一给定的排列方式计算的概率相乘。注意，后一步相当于使用了加法规则。

[例] 从一幅洗得很好的扑克牌中做了 3 次抽取，假定使用回置法，求至少得到 1 张 A 和一张 K 的概率是多少?

[解]按照题意，要在不同样本空间中考虑三种复合事件：抽到 1 张 A 和 1 张 K，另 1张非 A 非 K，用符号(AKO) 表示(其中“O”表示其他)；抽到 1 张 A 和 2 张 K，用符号(4KK)表示；抽到 2 张 A 和 1 张 K，用符号（AAK）表示。因为在不同样本空间中基本事件实现的概率不同，必须对它们加以区别。

次序为 AKO 的样本点实现的概率是

次序为 AKK 的样本点实现的概率是

次序为 AAK 的样本点实现的概率是

再考虑每个复合事件各含有多少种可能的排列方式

(AKK) 含有 3！ / 2!＝3 种排列方式

(AAK) 含有 3！ / 2!＝3 种排列方式

（AKO）含有 3！＝6 种排列方式

所以，在三次抽取中，至少得到 1 张 A 和 1 张 K 的概率是

[例] 假如对 1000 个大学生进行歌曲欣赏调查，发现其中有 500 个学生喜欢民族歌曲，400 个学生喜欢流行歌曲，而这些学生中有 100 人属于既喜欢民族歌曲又喜欢流行歌曲的，剩下来的学生两种歌曲都不喜欢。如果我们随机地从该总体中抽取一个学生，并设事件 A 为该学生喜欢民族歌曲，事件 B 为该学生喜欢流行歌曲。

①用数字证明 P(A 且 B)=P(A)P(B/A)=P(B)P(A/B)

②得到一个喜欢两种风格歌曲之一的学生的概率是多少？

③随机地选取一个由 3 个学生组成的样本，要求这三个学生全都有相同的欣赏方式，得到这种样本的概率是多少？

5. 运用概率方法进行统计推断的前提

简单随机抽样要求每一个个体拥有相同的被选入样本的机会。

严格来讲，由于我们实际上总是做不回置抽样，因此独立性的假定，是难以完全满足的。 只有在样本非常大，可以忽略。

一个随机样本具有以下的性质：不仅要给每一个个体以相等的被抽中的机会，而且要给每一种个体的组合以相等的被抽中的机会。

在要概括社区或其他空间上限定区域的单位的情况时，也必须注意到缺乏独立性的问题。

第三节　概率分布、期望值与变异数

1. 离散型随机变量的概率分布

离散型随机变量的取值是可数的，如果对 X 的每个可能取值 xi计算其实现的概率 Pi ，我们便得到了离散型随机变量的概率分布，即

2. 连续型随机变量的概率分布

连续型随机变量的取值充满某一区间，因而取某一数值讨论其概率是无意义的。为此，我们引进概率密度　　的概念来表达连续型随机变量的概率分布。

这样一来，随机变量 X 取值在区间{x1 ，x2}上的概率等于概率密度曲线　　下面 x1 与 x2 两点之间面积，即

所以

有概率密度的性质

3. 分布函数

为了从数学上能够统一对随机变量的概率进行研究引入分布函数　　的概念，它被定义为

有了分布函数，就可以很容易得到随机变量 X 取值在任意区间{x1 ，x2}上的概率，即

和　　(离散变量)或　　(连续变量)的关系，就像向上累计频率和频率的关系一样。不同之处在于，　　累计的是概率。但使用分布函数的好处是很明显的，它不仅在数学上统一了对离散型随机变量和连续型随机变量概率的研究，而且由于它计算概率的起点都固定为－∞，因而可以把概率值换算成表，以易于求得任何区间的概率，从而达到计算快捷和应用广泛之目的。

[例]　求两颗骰子点数的分布函数。

4. 数学期望

在前面统计分组的讨论中，我们在得到频数(或频率)分布后，为了对变量有系统概括的认识，分别研究了集中趋势和离中趋势。而对集中趋势和离中趋势量度，我们分别得到了平均指标和变异指标，其中最有代表性的是算术平均数和标准差。很显然，现在当我们面对随机变量的理论分布时，也要对随机变量的集中趋势和离中趋势作概括性的描述，这就引出数学期望和变异数这两个概念。

所谓数学期望，是反映随机变量 X 取值的集中趋势的理论均值(算术平均)，记作 E(X)。

[例]　一家保险公司在投保的 50 万元人寿保险的保单中，估计每 1000 保单每年有 15 个理赔，若每一保单每年的营运成本及利润的期望值为 200 元，试求每一保单的保费。

[解]　依题意知，利润的期望值

E(X)＝200(元)

设 x1 表示保费，x2 为理赔费[x2=-(500000- x1),]则可得

所以，x1＝7700(元)。即每一保单每年的保费应定在 7700 元。

数学期望也常常记为μ，在推论统计中同总体均值的记号，而　　则在推论统计中被作为样本均值的记号。数学期望和总体均值一样，都是唯一的，不过它是一个先验的理论值。由于它是用随机变量各取值分别乘以取值的概率来计算的，因此数学期望又可称为随机变量的加权算术平均数。样本均值依据统计数据计算而来，但它具有随机性。在统计推论中，E(X)　　，　　是“估计”。

5.　变异数

数学期望反映了随机变量的集中趋势，但仅知道集中趋势还不够，还应该知道随机变量在均值周围的离散程度，即离中趋势。变异数是综合反映随机变量取值分散程度的指标，其功能相当于描述统计中已讨论过的方差及标准差，记用 D(X)。

很显然随机变量 X 的变异数也可以写成

变异数的几个基本性质：

第七章 假设检验

我们在第一章就已经知道，推论统计有两个基本内容：①假设检验；②参数估计。有了概率和概率分布的知识，接下来我们要逐步掌握统计检验的一般步骤。既然按照数学规则得到的概率都不能用经验方法准确求得，于是，理论概率和经验得到的频率之间肯定存在某种差别，这就引出了实践检验理论的问题。随机变量的取值状态不同，其概率分布的形式也就不同。本章我们不仅要引出二项分布和正态分布这两个著名的概率分布，并且要将它们与抽样调查联系起来，以领会统计检验，并逐步拓宽其应用面。

第一节　二项分布

二项分布是从著名的贝努里试验中推导而来。所谓贝努里试验，是指只有两种可能结果的随机试验。在实际问题中，有许多随机现象只包含两个结果，如男与女，是与非，生与死，同意与不同意，赞成与反对等等。通常，我们把其中比较关注那个结果称为“成功”，另一个结果则称为“失败”。每当情况如同贝努里试验，是在相同的条件下重复 n 次，考虑的是“成功”的概率，且各次试验相互独立，就可利用与二项分布有关的统计检验。虽然许多分布较之二项分布更实用，但二项分布简单明了，况且其他概率分布的使用和计算逻辑与之相同。所以要理解统计检验以及它所涉及的许多新概念，人们几乎都乐意从二项分布的讨论入手。

1.　二项分布的数学形式

从掷硬币的试验入手。假定二项试验由重复抛掷 n 次硬币组成，已知硬币面朝上(成功)的概率是 p，面朝下(失败)的概率是 q（显然有 q＝1－p)。这样，对试验结果而

言，成功的次数（即硬币面朝上的次数）X 是一个离散型随机变量，它的可能取值是 0，1，2，3，…，n。而对 X 的一个具体取值 x 而言，根据乘法规则，我们立刻可以就试验结果计算出一种特定排列方式(先 x 次面朝上，而后 n－x 次面朝下)实现的概率，即

$$ppp\cdots pqqq\cdots q = p^x q^{n-x}$$

由于正确解决概率问题，光考虑乘法规则是不够的，还要考虑加法规则，于是就 x 次成功和（n－x）次失败这个宏观结果而言所包含的所有排列的方式数，用符号表示

这样，我们就得到了二项试验中随机变量 X 的概率分布，即

譬如，二项试验是将一枚硬币重复做 8 次抛掷，假设这枚硬币是无偏的，即 p＝q＝0.5，那么恰好得到 5 次面朝上的概率是

2. 二项分布讨论

③ $E(X)=\mu=np$， $D(X)=\sigma^2=npq$

④ 二项分布受 p 和 n 变化的影响，只要确定了 p 和 n，成功次数 X 的分布也随之确定。因此，二项分布还可简写作 $B(x;n,p)$

⑤二项分布的概率值除了根据公式直接进行计算外，还可查表求得。二项分布表的编制方法有两种：一种依据概率分布律 P(x) 编制(见附表 2)；另一种依据分布函数 F(x) 编制(见附表 3)。 其中

[例] 某特定社区人口的 10% 是少数民族，现随机抽取 6 人，问其中恰好 2 人是少数民族的概率是多少？

[解] 解法一：根据(7.3)式直接计算

解法二：根据附表 2 中纵列 n＝6 和横行 p＝0.1所对应 x 值，可直接查得 B(x；6，0.1)的概率值

$$B(2;\ 6,\ 0.1)=0.\ 0984$$

解法三：根据附表 3 求得

$$B(2;\ 6,\ 0.1)=F(2)-F(3)$$
$$=\ 0.1143-0.0159=0.0984$$

第二节 统计检验的基本步骤

二项分布是用数学或演绎推理的方法求得的一种理论分布。认识到概率分布是先验的理论分布这一点很重要，因为我们不禁要问，既然试验或抽样调查的结果仅与随机变量可能取值中的一个相联系，那么实际试验或样本调查对结果的概率分布及前提假设有没有一个检验的问题？具体来讲，对于一枚硬币被重复抛掷 8 次的二项试验，经验告诉我们，一共有 9 种可能的结果，而且实现这些结果的机会是大不相同的。研究者实际上从来不用经验的方法求得概率分布，因为通常我们只对一项试验进行一次或几次，抽取样本也是一个或至多不过几个。既然二项分布是按照数学规则得到的，那么对这 9 种结果的可能性我们应该作出何种评价呢?

如果实际试验（或抽样）得到的结果偏巧就是先验概率预示的最不可能出现的结果，那么我们是认定纯属巧合，还是开始对用数学或演绎推理方法求得的概率以及理想试验的种种前提假设产生怀疑?更准确地说，在一枚硬币被重复抛掷 8 次的这个二项试验中，究竟出现什么结果时，我们应该对二项分布及其前提假设产生怀疑呢?是不是只要不是得到 4 次成功 4 次失败这个最大可能性结果就开始怀疑，还是仅当出现 8 次成功或一次也不成功这两个极端情况时才产生怀疑呢?这就是统计检验的核心问题。

统计检验是指先建立一个关于总体情况的假设，继而抽取一个随机样本，然后以样本的统计量或者统计性质来检定假设。

大数定理表明：就大量观察而言，事件的发生具有一定的规律性。

根据概率的大小，人们处理的态度和方式很不一样。

在日常生活中，人们往往习惯于把概率很小的事件，当作一次观察中是极不可能看到的事件。例如，人们出门做事就有可能遇到不测事故，但却很少人因此而不敢出门。原因是：小概率事件极不可能发生。

1．建立假设

统计检验是将抽样结果和抽样分布相对照而作出判断的工作。取得抽样结果，依据描述性统计的方法就足够了。抽样分布则不然，它无法从资料中得到，非利用概率论不可。而不对待概括的总体和使用的抽样程序做某种必要的假设，这项工作将无法进行。比如通过掷硬币的实验得到二项分布，必须假设：①样本是随机的，试验中各次抛掷相互独立；②硬币是无偏的(或称是诚实的)，即 p＝q＝0．5。概括地说，必须首先就研究总体和抽样方案都做出假设，再加上概率论，我们就可以对各种可能结果做具体的概率陈述了。

2．求抽样分布

在做了必要的假设之后，我们就能用数学推理过程来求抽样分布了。比如在这一章开头，在硬币重复抛掷 n 次的理想实验中，我们计算了成功次数为 x 的宏观结果所具有的概率，得到二项分布。如果前提假设变动了，还可以求出其他形式的概率分布，如正态分布、泊松分布、卡方分布等等，它们都有特定的方程式。由于数学上已经取得的成果，实际上统计工作者要做的这项工作往往并不是真的去求抽样分布的数学形式，而是根据具体需要，确定特定问题的统计检验应该采用哪种分布的现成的数学用表。

3．选择显著性水平和否定域

在统计检验中，那些不大可能的结果称为否定域。如果这类结果真的发生了，我们将否定假设；反之就不否定假设。

在统计检验中，通常把被检验的那个假设称为零假设（用符号 H0 表示），并用它和其他备择假设(用符号 H1 表示)相对比。

在统计检验中，无论是拒绝或者接受原假设，都不可能做到百分之百的正确，都有一定的错误。第一类错误是，零假设 H0 实际上是正确的，却被否定了。第二类错误则是，H0 实际上是错的，却没有被否定。

遗憾的是，不管我们如何选择否定域，都不可能完全避免第一类错误和第二类错误，也不可能同时把犯两类错误的危险压缩到最小。对任何一个给定的检验而言，第一类错误的危险越小，第二类错误的概率就越大；反之亦然。一般来讲，不可能具体估计出第二类错误的概率值。第一类错误则不然，犯第一类错误的概率是否定域内各种结果的概率之和。

被我们事先选定的可以犯第一类错误的概率，叫做检验的显著性水平(用 α 表示)，它决定了否定域的大小。因此，有人也把第一类错误称之 α 错误 。相应地第二类错误被人称为 错误。

在原假设成立的条件下，统计检验中所规定的小概率标准一般取为 α =0.05或 α =0.01。

由 α 所决定的否定域与接受域之间的分界值被称为临界值， 如 Z α 。

如果抽样分布是连续的，否定域可以建立在想要建立的任何水平上，否定域的大小可以和显著性水平的要求一致起来（后面的正态检验就如此）。如果抽样分布是非连续的，就要用累计概率的方法找出一组构成否定域的结果。

根据否定域位置的不同，可以将假设检验分为双侧检验和单侧检验。

奈曼—皮尔逊 (Neyman —Pearson提出了一个原则 “在控制犯第一类错误的概率不超过指定值 的条件下, 尽量使犯第二类错误 小”按这种法则做出的检验称为“显著性检验”， 称为显著性水平或检验水平。

4．计算检验统计量

在完成了上述工作之后，接下来就是做一次与理想试验尽量相同的实际抽样(比如实际做一次重复抛掷硬币的试验)，并从获取的样本资料算出检验统计量。检验统计量是关于样本

的一个综合指标，但与我们后面参数估计中将要讨论的统计量有所不同，它不用作估测，而只用作检验。

5．判定

假设检验系指拒绝或保留零假设的判断，又称显著性检定。在选择否定域并计算检验统计量之后，我们完成最后一道手续，即根据试验或样本结果决定假设的取与舍。如果结果落在否定域内，我们将在已知犯第一类错误概率的条件下，否定零假设。反之，如果结果落在否定域外，则不否定零假设，与此同时，我们就有了犯第二类错误的危险。

[例] 若想通过抛掷 10 次硬币的实验来检验这个硬币无偏的零假设，通过双侧检验 0.10显著性水平，请指出否定域。如果单侧检验（p<0.5），又将如何？

[例] 某选区有选民 10000 人，其中属于工贸系统的有 4000 人，要产生代表 6 名。假定各系统选民都有同等机会当选代表，（1）代表是工贸系统人员的概率分布；（2）在 6 名代表中最可能是工贸系统人员占几名；（3）如果 6 名代表中有 4 名是工贸系统的人员，可以否定随机性的零假设吗？（α=0.05，单侧检验，p>0.4）

第三节 正态分布

如果说二项分布是离散型随机变量最具典型意义的概率分布，那么连续型随机变量最具典型意义的概率分布就是正态分布了。一般地讲，若影响某一变量的随机因素很多，而每个因素所起的作用不太大且相互独立，则这个变量服从正态分布。更为重要的是，不论总体是否服从正态分布，只要样本容量 n 足够大，样本平均数的抽样分布就趋于正态分布。

正态分布是最重要的概率分布：（1）许多自然现象和社会现象，都可用正态分布加以叙述；（2）当样本足够大时，都可用正态近似法解决变量的概率分布问题；（3）许多统计量的抽样分布呈正态分布。

（3）正态曲线的外形由σ值确定。对于固定的σ值，不同均值μ的正态曲线的外形完全相同，差别只在于曲线在横轴方向上整体平移了一个位置 。

2. 标准正态分布

Z 分数（标准正态变量）

用 Z 分数表达的标准正态分布，其概率密度为

3. 正态曲线下的面积

采用标准正态变量表达正态分布，使标准差得到了进一步阐明。我们看到，标准差是计算总体单位分布及其标志值变异范围的主要依据，下图说明了这一点。

（1）变量值在【　　μ-σ，μ+σ　　】之间的概率为 0.6826。

（2）变量值在【　μ-2σ，μ+2σ　】之间的概率为 0.9546。

（3）变量值在【　μ-3σ，μ+3σ　】之间的概率为 0.9973。

[例]　设随机变量 X 服从正态分布 N(168，12)，试求 P(X≤143)。

z 是负值，表示 X 的取值处于均值左边。由于曲线完全对称，所以使用正态分布表时可以忽略 z 的正负号。查表可知，正态曲线在均值与 z=2.08之间所含面积是 0.4812。由于总面积的一半是 0.5，因 P(X≤143)可以由下面计算求得

$$P(X\leqslant 143)=0.5-P(0\leqslant Z\leqslant 2.08)$$
$$=0.5-0.4812=1.88\%$$

这说明，X 的取值小于或等于 143 的概率大约是 2%。由于即将讨论的正态检验几乎都要涉及概率分布的尾端，所以此例说明的是一个非常普遍的问题。

4.　二项分布的正态近似法

通过前面的讨论，我们已经知道二项分布受成功事件概率 p 和重复次数 n 两个参数的影响，只要确定了 p 和 n，二项分布也随之确定了。 但是，二项分布的应用价值实际上受到了 n 的很大限制。也就是说，只有当 n 较小时，我们才能比较方便地计算二项分布。所幸的是，二项分布是以正态分布为极限的。所以当 n 很大时，只要 p 或 q 不近于零，我们就可以用正态近似来解决二项分布的计算问题。即以 $np=\mu$、$npq=\sigma^2$，将 B(x；n，p)视为 N(np，npq)进行计算。在社会统计中，当 n ≥30，np、nq均不小于 5 时，对二项分布作正态近似是可靠的。

第四节 中心极限定理

一旦统计的学习进入到推论统计，我们就必须同时与三种不同的分布概念打交道，即总体分布、样本分布、抽样分布。为了不产生混淆，视分布不同，将统计指标的符号加以区别是完全必要的。对那些反映标志值集中趋势和离中趋势的综合指标，尤其对均值和标准差(或方差)。

1．中心极限定理

我们知道，概率论中用来阐明大量随机现象平均结果的稳定性的定理，是著名的大数定理。其具体内容是：频率稳定于概率，平均值稳定于期望值。但是，大量随机现象的稳定性不仅表现在平均结果上，同时也表现在分布上，这就是中心极限定理所要阐明的内容。显然，推论统计需要有一座能够架通抽样调查和抽样分布的桥梁。中心极限定理告诉我们：如果从任何一个具有均值 μ 和方差 σ^2 的总体(可以具有任何分布形式)中重复抽取容量为 n 的随机样本，那么当 n 变得很大时，样本均值的抽样分布接近正态，并具有均值 μ 和方差　　。