中国社会学函授大学教材之

社會研究中的統計分析

（下册）

（函授教材，不得翻印）

中国社会学函授大学编印

目　录

（下　册）

第五章　从局部推论到总体……（1）
第一节　统计推论的概念……（1）
第二节　总体参数的点估计……（5）
第三节　总体参数的区间估计……（14）
第六章　统计假设检验……（29）
第一节　统计假设……（29）
第二节　假设检验的基本概念……（33）
第三节　均值检验……（38）
第七章　相关与回归……（47）
第一节　散布图……（47）
第二节　相关系数……（51）
第三节　回归直线……（55）
第四节　利用回归方程进行预测……（58）
第八章　列联表分析……（63）
第一节　列联表……（63）
第二节　列联表检验……（68）
第三节　列联强度……（77）
第九章　研究报告的书写……（81）

第五章　从局部推论到总体

第一节　统计推论的概念

一、什么是统计推论

所谓统计推论就是根据局部资料（样本），对总体的性质进行推断和估计。它是和抽样调查紧密连系在一起的。在第二章的描述统计分析中，我们也谈到如何将调查到的资料进行整理，制成表和图，以便进行分析。但它的处理是仅就调查的结果而言的。也就是说，如果调查100名工人的平均结婚年龄，算出来的平均结婚年龄 $\bar{X}$ 和标准差σ都是仅就这100名工人所具有的特征，它不具有外推的性质。但是我们知道，抽样调查的意义就在于通过局部的研究要能把结果外推到整个研究的总体。为此，我们曾经谈过样本需要有代表性。但是不管样本具有多好的代表性，仍然免不了存在抽样误差。因此通过样本计算的特征值，如果外推到更大的总体的话，还必须经过必要的数学处理，这些就是统计推论的内容。

二、“抽样结果与总体参数不一样”的实例。

为了对抽样结果与为之要推论的总体参数两者不完全一样有一个形象的了解，下面我们来作一次实验。

某工厂共有100名工人，其中男性占50名，女性占50名。设某研究人员事先并不知道工厂的性别比。为此，他希望通过抽查10名工人，统计他们的性别比来推论全厂工人性别比。为了满足随机抽样的要求，他把100名工人作成100个阄，充分搅匀，从中抽取10个，看看抽出的结果，其中有几名女性，几名男性。但是由于我们作实验的目的，对具体抽到人的姓名并不感兴趣，我们

只需要知道是男性或是女性。因此，我们也可以简化为棋子来作此实验。例如用围棋子100个，其中黑、白两色各50个。并设白色棋子代表女性，黑色棋子代表男性。然后把这100个棋子放入袋中，充分摇匀。从中抽取10枚。下面是反覆抽取10次(10样本)的结果：

样本	白棋	黑棋
1	6	4
2	6	4
3	5	5
4	4	6
5	4	6
6	4	6
7	5	5
8	6	4
9	6	4
10	4	6

如果按照白棋/黑棋出现次数进行统计：

白棋/黑棋	6/4	5/5	4/6
样本出现次数	4	2	4

可见，在10次样本中，真正出现白棋/黑棋为5/5的只有2次，而其它8次不是白棋出现个数多于黑棋，就是黑棋个数多于白棋。实际上，不仅会出现4/6或6/4的情况，其它3/7；2/8；1/9；0/10；……等等都是可能出现的。读者有兴趣的话，也可以

按上述方法实验一下，看看结果如何。前面我们说过，我们是把白棋代表为女性，黑棋代表为男性。总数100枚代表全厂工人总数为100人。其中黑色棋子50枚表示全厂男性工人总数为50名。白色棋子50枚表示全厂女性工人总数为50名。可见，在工厂性别比为50/50的情况下，上面的实验中，实际只出现两次是5/5。而实际工作中，我们只进行一次抽样，也就是只有一个样本，因此很难说，我们碰到的是哪次结果。因此，样本中的性别比很难直接代表总体中的性别比。这也是从抽样结果推论到总体时的难点所在。因此，如果有人根据一次抽样问卷调查，发现男性有47%喜欢收看体育比赛的电视节目，女性有53%喜欢收看体育比赛的电视节目，就作出女性比性男更喜欢收看体育比赛的电视节目，这样的结论就欠把握了。因为，在一个收看体育比赛没有性别差的总体里，一次抽样的结果，是很难正好碰上是50%：50%的。很可能是既有男性比女性多一些的情况也有女性比男性多一些的情况。

那么，如果要问为什么会出现抽样结果不同于母体的情况呢？原因就在于我们所研究的现象具有随机性，也就是多种情况性。例如性别有男、女两种情况。电视节目有收看或不收看两种情况。因此它就不会象抽查一滴水的成份是H_2O，就可以推论到所有的大洋、大海 其水的成份都是H_2O那样简单了。这一基本概念是我们作抽样调查时，必须首先明确的：只要资料处理的结果，须要外推到没有调查到的更大范围（全体）就必须要用统计推论的知识，去检验它的推论是否合理、可靠。

三、统计推论内容简介

统计推论的内容，简单说有两大部分。第一部分是通过抽样数据如何对总体的参数进行估计。例如通过北京市1000户的家计调查，估计全市人民的平均收入，即平均值的估计。也可以对全市拥有电视机的户数进行百分比的估计，也就是百分比或成数的估计。也可以对某一个号数，例如均值、百分比分散程度进行估计，即标准差的估计等等。这些参数的估计，细分起来又可作两

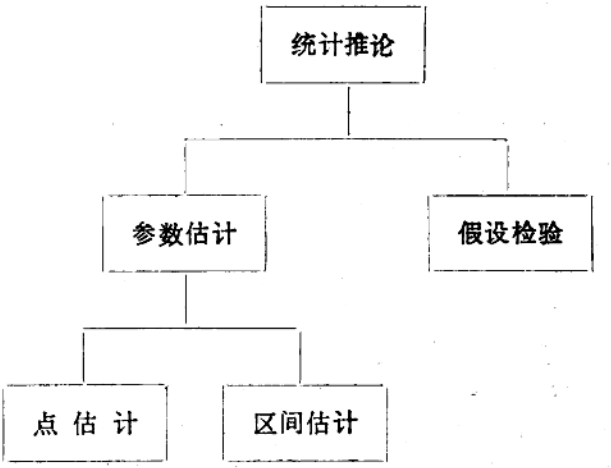
类。一类只简单的给出一个估计值。例如，估计的结果，北京市人平均月收入为 $\bar{X}=80$元。或拥有电视机的户数占全市60％等等，这一类简称参数的点估计。另一类，也是更科学的估计方法。它指出估计的范围，以及它的可靠性有多大。例如，估计的结果为有95％的把握说：北京市的人平均月收入在100元—60元之间。或有95％的把握说：北京市拥有电视机的户数占总户数的50％—60％之间。这类估计的特点是给出估计的范围，并指示估计范围可靠的程度，简称区间估计。这一部分统称参数估计问题。

另一部分问题是属于假设检验问题。它的内容是如何调动抽样数据对抽象层次的假设进行检验。实际上，我们在第一章中所介绍的社会调查方法，其中心思想就贯穿了统计检验的全部过程。因此，可以说，由于数理统计的应用，使得社会科学的研究与自然科学的研究，建立在同等的研究层次之上了。在社会调查方法中，我们谈到，通过社会现象抽象层次的观察，建立了一定的概念、命题或理论。经过操作化定义的测量，建立起若干定量化的假设，又称原假设H_0，

$$H_0:\quad P_0=P_1$$

（例如，它可以表示是甲、乙两队的平均收入相等。或甲、乙两队赞成某项政策的百分比相等等等）。

下一步的工作，将是实地进行调查，例如我们用抽样的方法进行调查。根据本节二所介绍，我们知道，既或两队总体确实存在$P_0=P_1$。抽样的结果，也可能出现调查的结果是$P_0\neq P_1$。而统计学的作用，就是对样本数据按一定的程序进行分析计算，根据计算的结果，对总体假设作出合理的推断。并对这种推断（例如：接受原假设$P_0=P_1$ 或拒绝原假设$P_0\neq P_0$）可能发生错误的概率，预先加以确定。这就是统计推论的思路。总结起来说，统计推论的内容可以用以下的图来表示：



第二节　总体参数的点估计

在具体介绍统计推论前，需要说明的是，考虑到本书的读者都是从事具体工作的，因此下面内容的介绍，尽量避免数学上的解释，多用浅显易懂的实际例子来说明。这样的结果，虽然从数学的角度来说可能不够准确或不够严谨，但能使读者可以更形象地理解它。其次，统计推论从它的数学基础来说，必竟还是比较深奥的，它是以概率论为基础的统计推论方法。因此，当读者对它的数学推导略去不管之后，最主要的是学习它的结论。但是任何一个结论、一个公式都不是万能的，都是针对某一种前提所得出的，我们称之谓假定。假定是讨论问题的前提，是得出公式的条件。因此，在使用任何一个具体公式之前，必须先了解它的前提（假定），然后，再进行运用。因此，本书重在交待各计算公式的假定，这点是学习今后诸章内容必须注意的。

一、总体均值的点估计

前面讲过，所谓参数的估计，用通俗的话来说，就是根据抽

样结果来合理地、科学地猜一猜，全体的参数大概是什么？或在什么范围？参数估计的问题是随时都可以碰到的。从日常生活直到科学研究中都会碰到它。比如说，为了决定星期日是否带孩子去逛公园，需要对天气有一个估计。一种新药是否投产，有赖于通过试验阶段的有效率进行估计。商店进货的挡次，需要对当地居民平均收入有所估计。社会学家需要对家庭的平均人口，平均子女数等等有所估计。而参数的点估计则是用样本中计算出来的一个参数来估计总体参数。具体说，如果总体参数指的是总体的均值，则是总体均值的点估计。

有那些问题属于总体均值的点估计内容呢？

首先，调查的结果必须是随机抽样得到的。也就是说，为了对一个全体的均值进行正确的估计，抽样必须保证是随机的。否则样本将失去代表性。例如，你想了解一个学校的平均工资，如果你只调查教授、领导干部的收入，那么，用这样的抽样调查算出的平均工资，显然比实际全校的平均工资要高。反之，如果抽取的调查对象，全是低收入的勤杂工，用这样的平均工资来代表全校的平均工资，显然又会偏低。为了排除主观干扰，最好是先有一份全校教职、员、工的花名册，然后抽到谁就调查谁，以此来保证随机性。这点很重要，如果随机性不能保证，下面一切估计将无任何可靠性可言。同时，由于抽样的随机性，属于讨论问题的前提或假定，所以，如果它不能保证的话，后面一切数学计算也是检验不出来的。就象产品的质量评比，如果厂家只挑几个精雕细刻的产品，拿去质量评比，其结果绝不能反映它整个产品的质量一样。

其次，要构成一个总体均值的估计问题，还必须要求调查的结果有量的大小。如果问卷中的回答，属于“是”或“不是”、“同意”或“不同意”、“男性”或“女性”一类，那么，对于这样的总体估计，一般不会是总体均值的估计问题。而只有问题的回答有量的大小，例如“25岁”、“3个孩子”、“10年”……

等等，这类数据，才有可能算出均值。因此，变量的层次，严格的说必须是定距的或定比的。对于定序变量，只能是近似的运用。而定类变量，例如男、女，如果我们把男＝0、女＝1，然后，求其均值，那将毫无实际意义。因此，问题是否属于均值估计问题，还必须注意收集资料的性质，只有资料具有量的概念，使用平均值才有实际意义。

在以上两点前提下，那么，通过抽样调查得到如下几个观察值（或调查值）

X_1，X_2，X_3……Xn

就可以用几个观察值的平均值

$$X = \frac{1}{n}\sum_{i=1}^{n} X_i$$

作为总体均值的点估计值。

$\sum_{i=1}^{n} X_i$—是$X_1 + X_2 \cdots + X_n$的缩写方法。

Σ—读作“西格吗”，表示总和。

$i=1$—表示总和的第一个值从X_1加起。

n—表示总和的最后一个值是X_n。

例1：为了解某区家庭平均子女数，作了10户的随机抽查，结果有

3；2；2；1；4；2；1；1；1；2。求该区家庭平均子女数的估计值。

〔解〕：$X = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{10}(3+2+2+1+4+2+1+1+1+2)$

$= 1.9$

例2：为了解某工厂女青年的平均结婚年龄，作以下5名的随

机抽查，结果有

24；23；26；32；17。

求该工厂女青年的平均结婚年龄估计值。

〔解〕：$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i=\frac{1}{5}(24+23+26+32+17)$

$=24.4$(岁)

例3：某研究所为了解女科技人员的家务负担问题，对20名女科技人员调查了每天用于家务劳动的时间（小时）：

3；5；4；6；4；2；4；3；5；6；

1；4；3；4；5；7；3；2；5；3。

求该研究所女科技人员每天平均家务劳动的估计值。

〔解〕：$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i=\frac{1}{20}(3+5+4+6+4+2+4+3+5+6+1$

$+4+3+4+5+7+3+2+5+3)$

$=3.95$(小时)

通过以上的例子，可以看出，我们具体运算的都是样本观察值的平均值。例如$\overline{X}=1.9$(个)；$\overline{X}=24.4$(岁)；$\overline{X}=3.95$(小时)。但是，统计推论的涵义，却在于告诉我们，这些样本的平均值是总体（例如某区、某工厂、某研究所）平均值的一个很好的估计值。

读者可能会产生这样的问题，如果对于总体均值的估计，确如前面例子所示，那再好不过了。我们对于一个很大的总体，例如一厂区、一个工厂、一个科研单位，只需抽查极少的几个（例如10名、5名、或20名）就可以对总体均值进行估计了。但是，实际上，数理统计告诉我们，抽样的个数（又称样本的容量）越大，对总体参数估计的准确性也越大。这就是为什么对于一个大的总体，例如一个学校，一个工厂抽样调查的人数，一般至少要几百名甚至1000名以上的缘故。

下面总结一下，总体均值的点估计：

总体均值的点估计：

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

X_i——样本的观察值。

假定：简单随机抽样

数据具有定量的特征。

二、总体成数（比率）的点估计

以上谈了总体均值的点估计，其中强调了数据必须具有定量的特征，如果数据不具有定量的特征，例如只有类别之分，是定类型数据又将如何研究它呢？对于定类型数据，由于我们只能数它属于某一类的有多少个。因此，我们只能研究某一类的事或物，占总数中的比率、百分比或成数是多少。反之，如果总体中某一类所占的成数，我们不知道，而是希望通过抽样的结果，看看某一类在抽样总数中所占的比例，去估计总体中这一类所占的比例，那么这类问题，很可能就是总体成数的估计问题了。如果估计值只需要单一的成数，那就是总体成数的点估计问题了。

对于总体成数的估计问题，同样须要抽样满足随机性，这里指的是简单随机抽样。在这样的前提下，如果抽样调查 n个，其中有m个属于要估计的这一类，那么

$$f = \frac{m}{n}$$

将作为总体中研究的这一类成数的点估计值。

例1：为了解某村计划生育的态度，对100名已婚者进行了抽样调查，发现其中有50名领取了独生子女证。求该村赞成独生子女比例的估计值。

〔解〕根据样本中100名调查，其中赞成独生子女的比例为：

$$f = \frac{50}{100} = 0.5$$

因此，可用样本的比例f＝0.5，作为全村赞成独生子女比例的估计值。

例2：工会为了解春游期间要租用几辆公共汽车，在全厂10000名职工中进行了共100人的随机抽样调查。统计结果，其中有20名愿意出外春游。设每辆公共汽车可载50名乘客，问估计要预租多少辆公共汽车。

〔解〕根据抽样调查，愿意外出春游的比率为：

$$f = \frac{20}{100} = 0.2$$

我们可以用f作为全厂将参加外出春游比率的点估计值，因此全厂将有

$$10000 \times 0.2 = 2000 \text{（人）}$$

参加春游。又因每辆公共汽车可容乘客50人，因此有：

$$\frac{2000}{50} = 40\text{(辆)}$$

即：估计预租40(辆)，可满足全厂春游的需要。

在成数的点估计中，如同总体均值的估计值一样。调查的人数越多，样本容量n越大，成数的点估计值接近真正总体成数的可能性越大。下面总结一下，总体成数的点估计：

$$f = \frac{m}{n}$$

n—观察总次数

m—n次观察中，所研究的A事件共出现的次数。

假定：简单随机抽样

数据仅具有分类的特征。

三、总体方差（及标准差）的点估计

除了对总体反映集中趋势的均值，成数须要进行估计外，在很多情况下，还要对总体的分散情况进行估计。在第一章里，我们曾介绍过，描述总体分散情况的数量特征，可用方差（及标准差）来表示。

方差：$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2$

标准差：$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N}}$

它表示观察值距离均值的平均离散程度。

如果现在不知道总体的分散情况，而是希望通过一个样本的观察值：

$$X_1, X_2 \cdots\cdots X_n$$

来对总体的方差σ^2进行估计。那么，总体方差的点估计值为：

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

比较S^2（我们称作为样本方差）和σ^2（我们称作为总体方差），可以发现，两者唯一的区别是，S^2中分母是“$n-1$”，而σ^2中分母是N。也就是说，如果进行的是全面调查，例如全工厂共1000人。根据1000人的工资，求其工资的方差σ^2，这时分母是用1000来除的。但如果从1000人中抽查一小部分，例如50名，然后用这被抽查到的50名职工的工资，求其方差（样本方差），并借此作为总体工厂1000人工资方差σ^2的估计值，那么，分母应用$50-1$来除。当然，当抽样人数比较多时，分母用n来除，或是用$n-1$来除，差别就很小了。这时，两个公式都可以用。

同时，对于方差σ^2估计公式，也存在n越大，估计准确的可能性也越大。

根据方差点估计公式S^2，可以看出，当观察值很大时，例如，每一个观察值都是几百或几千，那么几个这样的数加起来，数值就更大，下面结合实例，谈谈如何进行简化运算。

例：根据抽样调查，共有以下12个观察值：

232.50　232.48　232.15　232.53　232.45

232.48　232.05　232.45　232.60　232.30

232.30　232.47

求其总体的均值及方差。

〔解〕根据点估计公式：

$$\overline{X}=\frac{1}{12}\sum_{i=1}^{12}X_i$$

$$S^2=\frac{1}{12-1}\sum_{i=1}^{12}(X_i-\overline{X})^2$$

分别是总体均值与方差的点估计。

因此有：$\overline{X}=\frac{1}{12}(232.50+232.48+\cdots\cdots 232.47)$

$$=\frac{1}{12}\times 2788.76=232.3967$$

$$S^2=\frac{1}{12-1}\sum_{i=1}^{12}(X_i-\overline{X})^2$$

$$=\frac{1}{12-1}\left(\sum_{i=1}^{12}X_i^2-12\times(\overline{X})^2\right)$$

$$=\frac{1}{11}(232.50^2+232.48^2+\cdots\cdots+232.47^2-12\times$$

$$\times 232.3967^2)$$

$$=\frac{1}{11}\times 0.29445=0.02677$$

根据以上的计算，可以看出，由于每一个观察值都很大，因此，计算起来还是很繁琐的。但是，我们如果细心观察，就会发现，每一个观察值中都含有232，因此，如果把每一个观察值，在计算均值之前，都预先减去232的话，那么，新的观察值就小得多了：

$$X'_i = X_i - 232$$

X'_i: 0.50 0.48 0.15 0.53 0.45 0.48 0.05 0.45
0.60 0.30 0.30 0.47

$$\bar{X}' = \frac{1}{12}\sum_{i=1}^{12} X_i' = \frac{1}{12}(0.50 + 0.48 + \cdots\cdots + 0.47)$$

$$= \frac{1}{12} \times 4.76$$

$$= 0.3967$$

但是，最后别忘了我们要求的均值是$\bar{X}$而不是$\bar{X}'$。因此，最终的答案将是

$$\bar{X} = 232 + \bar{X}'$$

$$= 232.3967$$

同时由于X'_i是X_i减去常量232，而常量是不变、没有方差的。因此，计算样本方差S^2，用数据X_1，X_2，……X_{12}

或用数据X_1'，X_2'……X_{12}'

都是一样的。现在我们用X_i'来计算

$$S^2 = \frac{1}{11}\sum_{i=1}^{12} (X_i'^2 - 12\,\bar{X}')^2$$

$$= \frac{1}{11}(0.50^2 + 0.48^2 + \cdots\cdots 0.47^2 - 12 \times 0.3967^2)$$

$$= \frac{1}{11}(2.1826 - 1.8881) = 0.02677$$

不难看出，后面的计算方法比前者简化。

第三节　总体参数的区间估计

前面谈了用样本观察值计算出来的一个点值，作为总体均值或方差的估计值。但是，对这样的估计方法，我们还远远不能满足。因为，真正总体的参数（均值或方差）我们并不知道，而一次样本算出的均值和方差，我们又无法说出有多大的可能，距离真正的参数有多远。还以第一节中的例子来说，袋中虽然有50枚白色棋子，50枚黑色棋子，当随机从中摸10枚时，既可能出现5枚白棋子，5枚黑棋子，也可能摸出的是白棋子数目多于黑棋子数目，也可能是黑棋子数目多于白棋子数目，甚至最极端情况下，摸出10枚可以全是黑色棋子或全是白色棋子。而当我们只作一次抽样，而且出现的情况，又正是这种最极端情况时，如果用这样的样本值去估计总体参数，那相差就太远了。而且，最困难的是由于我们对总体情况不知道，因此我们并不知道是否估错，或估错的可能性有多大。下面介绍的区间估计方法，则是对未知参数作一区间范围的估计，同样还要指出被估计参数包含在这一区间范围的可能性有多大。

对于区间估计，同样要保证样本的抽取是随机的，排除主观因素干扰的。例如由调查员或单位负责人来指定被访者都将是错误的。此外，下面介绍的方法属于大样本方法，也就是样本容量或称调查的对象一定要超出50名以上$n \geq 50$，否则，所用的公式将是不正确的。根据调查数据的性质是否具有量的概念，正如前节中所介绍的点估计那样，可以分作总体均值的区间估计和总体成数的区间估计两类。以下分别介绍它们的内容。

一、总体均值的区间估计

总体均值的区间估计，我们写成（Q_1，Q_2）的形式，它表示总体的参数Q估计在Q_1到Q_2之间。例如，我们估计，北京市区女青年的平均初婚年龄在23.5（岁）——24.5岁之间。这时，

就可写作〔23.5（岁），24.5（岁）〕。这样的区间也可写成〕24岁－0.5岁，24岁＋0.5岁〕。它表示初婚年龄在24岁左右，上、下偏离约为0.5岁。或者索性写成24岁±0.5岁。从实际意义来说，它们表示的内容都是相同的。我们把Q_2-Q_1称作估计区间的宽度。在数学上称作置信区间。例如，上面所举的北京市女青年的平均初婚年龄估计宽度为

$$24.5岁-23.5岁=1岁$$

估计区间的宽度Q_2-Q_1与估计的可信度是有连系的。所谓可信度就是估计的可靠性，在数学上称置信度。估计宽度的区间愈宽，Q_2-Q_1值愈大，则可靠性也就越大。这在日常生活中，是很容易体会到的。例如，有人让你估计班级某次考试平均分。如果你的估计是0—100分，那你永远说对了，因为这样的估计有100%的把握，可靠性为100%。相反，如果你的估计愈窄，例如60—70分，或更窄的区间63—65分等等，那么，估计错的可能性就愈变愈大。可见，估计的宽度与可靠性成正比。可靠性我们用P表示，在数学上称作概率。P的最大值为1，它表示有100%的可靠性或100%的把握。P的最小值为0，它表示完全没有可能。一般情况下，P的取值在0—1之间：

$$0\leq P\leq 1$$

那么，能否把估计区间的宽度尽量增大些，这样估计的可靠性也随之增加，岂不两全其美吗？实际上，不能这样做，因为估计区间越宽，意味着它的估计精确度愈低。实事上，如果你对班次估计的平均分数是0—100分的话，别人一定会觉得这是一句废话，因为，它虽然有100%的估计可靠性，但却没有提供任何信息。可见，估计可靠性的增加，会伴随着估计精确性的下降，两者是相互制约的。

那么，具体说，估计的区间与可信度有什么关系呢？对于大样本来说，总体均值的区间估计为：