第5章 参数估计

在实际应用中，人们常常需要根据手中的数据分析或推断全部数据的本质，即根据样本推断总体的分布或数字特征等，这需要统计推断的相关知识。所谓统计推断是指根据样本对总体分布或分布的数字特征等作出合理的推断。统计推断是数理统计学的一个重要分支。

本章介绍的参数估计是统计推断的重要内容之一。

5.1 参数估计概述

参数估计是用样本统计量估计总体参数。例如，用样本均值 $\bar{x}$ 估计总体均值 μ，用样本方差 s^2 估计总体方差 σ^2，等等。如果用 θ 表示总体参数，参数估计也就是如何用样本统计量来估计总体参数 θ。

用来估计总体参数的统计量称为估计量，用符号 $\hat{\theta}$ 表示。样本均值、样本比例、样本方差等都可以是一个估计量。用来估计总体参数时计算出来的估计量的具体数值称为估计值。例如，要估计一个地区的平均收入，不太可能将该地区的所有人的收入都统计一遍。一般的做法是：选择一部分人员进行收入统计，这部分被选择人员的收入数据就是样本数据。通过样本数据可以求出样本均值 $\bar{x}$，它就是一个估计量 $\hat{\theta}$。假设计算出的样本均值为10 000，则这个10 000就是估计值。一般而言，参数估计可以分为点估计和区间估计两种。

5.1.1 点估计

点估计是依据样本估计总体分布中所含的未知参数或未知参数的函数，通常它们是总体的某个特征值，如期望、方差等。点估计问题就是要构造一个只依赖于样本的量，作为总体的估计量。在用点估计值代表总体参数值的同时，还必须给出点估计值的可靠性，也就是说，必须能说明点估计值与总体参数值的接近程度。

5.1.2 区间估计

区间估计是利用抽取的样本，根据一定的正确度与精确度的要求，构造出适当的区间，作为总体分布的未知参数的值所在范围的估计。与点估计不同，进行区间估计时，根

据样本统计量的抽样分布可以对样本统计量与总体参数的接近程度给出一个概率度量。在区间估计中，由样本统计量所构造的总体参数的估计区间称为置信区间。其中，区间的最小值称为置信下限，用 θ_1 表示；区间的最大值称为置信上限，用 θ_2 表示。总体参数落在 $[\theta_1,\theta_2]$ 区间的概率为 $p=1-\alpha$，称为置信水平，α 称为显著性水平。置信区间表明总体参数的误差范围，置信水平表明总体参数落在置信区间的概率。

5.1.3 估计量的评价标准

在参数估计时，可以构造很多估计量，但不是每一个估计量都一样优良。通常评价估计量的标准有无偏性、有效性和一致性。

(1) 无偏性。无偏估计量的抽样分布的数学期望等于被估计的总体参数。设总体参数为 θ，所选估计量为 $\hat{\theta}$，如果 $E(\hat{\theta})=\theta$，则称 $\hat{\theta}$ 为 θ 的无偏估计量。

(2) 有效性。一个无偏的估计量并不意味着就非常接近被估计的参数，它还必须符合与总体参数的离散程度充分小这一标准。同一总体参数的两个无偏估计量中标准差更小的有效性更高。

(3) 一致性。指随着样本量的增大，点估计量的值越来越接近被估计的总体参数。换言之，一个大样本给出的估计量要比一个小样本给出的估计量更接近总体参数。

5.2 总体均值区间估计

总体均值区间估计是一种常见的参数估计问题，例如，根据抽样调查结果对某公司的平均利润率进行估计，根据参与调查学生的每月平均生活消费估计在校大学生月平均生活费等。总体参数的估计区间通常是由样本统计量加减抽样误差而得到的。根据样本统计量的抽样分布，能够对样本统计量与总体参数的接近程度给出一个概率度量。

在对总体均值进行区间估计时，需要考虑总体是否为正态分布、总体方差是否已知、用于构造估计量的样本是大样本还是小样本等情况。

5.2.1 总体方差已知或大样本情况下的估计

当总体服从均值为 μ、方差为 σ^2 的正态分布时，取自该总体的随机样本 $x_1,x_2,\cdots,x_n$ 的样本均值 $\bar{x}$ 应服从均值为 μ、方差为 σ^2/n 的正态分布，即 $\bar{x}\sim N\left(\mu,\dfrac{\sigma^2}{n}\right)$。对 $\bar{x}$ 进行标准化，对应的 z 值为 $z=\dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$。在给定显著性水平 α 或置信水平 $p=1-\alpha$ 的条件下，有

$$p(-z_{\alpha/2}<z<z_{\alpha/2})=1-\alpha$$

在总体方差 σ^2 已知的情况下，总体均值 μ 的置信区间为

$$\left[\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

在总体方差未知但为大样本(样本容量 $n \geqslant 30$)的情况下,可以用正态分布进行近似计算,使用样本方差 s^2 来代替总体方差 σ^2。此时总体均值 μ 的置信区间为

$$\left[\bar{x}-z_{\alpha/2}\frac{s}{\sqrt{n}},\bar{x}+z_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

例 5.1 某学校需要了解近几年计算机专业毕业生的月工资情况,现随机抽取了 800 名计算机专业毕业生,通过计算得到他们的平均月工资为 1800 元,总体标准差为 39 元。在 95%的置信水平下,求计算机专业毕业生的平均月工资置信区间。

步骤如下:

(1) 打开“第 5 章参数估计.xlsx”文件,选择“例 5.1”工作表,在 A1:B4 单元格区域中输入本例给出的数据。

(2) 在 A6 单元格中输入“标准误差”,在 B6 单元格中输入公式“=B2/SQRT(B3)”。

(3) 在 A7 单元格中输入“Z 值”,在 B7 单元格中输入公式“=ABS(NORM.S.INV.S((1-B4)/2))”。

(4) 在 A9 单元格中输入“极限误差”,在 B9 单元格中输入公式“=B6 * B7”。

(5) 在 A11 单元格中输入“置信上限”,在 B11 单元格中输入公式“=B1+B9”。在 A12 单元格中输入“置信下限”,在 B12 单元格中输入公式“=B1-B9”。

最终结果如图 5.1 所示。

	A	B
1	样本均值	1800
2	标准差	39
3	样本容量	800
4	置信水平	0.95
5		
6	标准误差	1.378858
7	Z值	1.959964
8		
9	极限误差	2.702512
10		
11	置信上限	1802.703
12	置信下限	1797.297

图 5.1 例 5.1 最终结果

例 5.2 某快递公司要分析第一季度日均快递件数,为此从第一季度抽取了 40 天的快递件数,如图 5.2 所示。在 95%的置信水平下,求该快递公司第一季度日均快递件数的置信区间。

	A	B	C	D	E	F	G	H	I	J
1	890	873	679	982	1031	760	1130	965	1210	890
2	1502	1210	730	1130	965	977	1320	1452	899	1502
3	1298	899	989	1320	1452	947	847	1109	980	808
4	793	709	823	989	1190	976	1120	1010	1121	1008

图 5.2 从第一季度抽取的 40 天的快递件数

步骤如下:

(1) 打开“第 5 章参数估计.xlsx”文件,选择“例 5.2”工作表,在 A1:J4 单元格区域中输入图 5.2 中的数据。

(2) 计算样本均值。在 A6 单元格中输入“样本均值”,在 B6 单元格中输入公式“=AVERAGE(A1:J4)”。

(3) 计算标准差。在 A7 单元格中输入“标准差”,在 B7 单元格中输入公式“=STDEV.S(A1:J4)”。

(4) 计算样本容量。在 A8 单元格中输入“样本容量”,在 B8 单元格中输入公式“=COUNT(A1:J4)”。

(5) 在 A9 单元格中输入“置信水平”，在 B9 单元格中输入 0.95。

(6) 计算标准误差。在 A11 单元格中输入“标准误差”，在 B11 单元格中输入公式“=B7/SQRT(B8)”。

(7) 计算 Z 值。在 A12 单元格中输入“Z 值”，在 B12 单元格中输入公式“=ABS(NORM.S.INV((1−B9)/2))”。

(8) 计算极限误差。在 A14 单元格中输入“极限误差”，在 B14 单元格中输入公式“=B11 * B12”。

(9) 计算置信上限与置信下限。在 A16 单元格中输入“置信上限”，在 B16 单元格中输入公式“=B6+B14”。在 A17 单元格中输入“置信下限”，在 B17 单元格中输入公式“=B6−B14”。

最终结果如图 5.3 所示。

	A	B	C	D	E	F	G	H	I	J
1	890	873	679	982	1031	760	1130	965	1210	890
2	1502	1210	730	1130	965	977	1320	1452	899	1502
3	1298	899	989	1320	1452	947	847	1109	980	808
4	793	709	823	989	1190	976	1120	1010	1121	1008
5										
6	样本均值	1037.125								
7	标准差	219.0873								
8	样本容量	40								
9	置信水平	0.95								
10										
11	标准误差	34.64074								
12	Z值	1.959964								
13										
14	极限误差	67.89461								
15										
16	置信上限	1105.02								
17	置信下限	969.2304								

图 5.3　例 5.2 最终结果

由上述两个例子可以发现，计算总体方差已知或大样本情况下的均值置信区间的步骤几乎是一样的，可概括如下：

(1) 利用给定的数据计算样本均值、标准差、样本容量和置信水平。

(2) 根据步骤(1)中的标准差和样本容量计算标准误差，计算公式为

标准误差 = 标准差 /SQRT(样本容量)

(3) 根据步骤(1)中的置信水平计算 Z 值，计算公式为

Z 值 = ABS(NORM.S.INV((1 − 置信水平)/2))

(4) 根据步骤(2)求出的标准误差和步骤(3)求出的 Z 值计算极限误差，计算公式为

极限误差 = 标准误差 × Z 值

(5) 根据步骤(1)中的样本均值和步骤(4)求出的极限误差计算置信区间，计算公式为

置信上限 = 样本均值 + 极限误差

置信下限 = 样本均值 − 极限误差

5.2.2 总体方差未知且为小样本情况下的估计

小样本是指样本容量 $n<30$ 的情况。本节讨论的情况的前提为总体方差未知并且样本容量小于 30。在这种情况下，可以使用样本标准差 s 代替未知的总体标准差，新的统计量服从自由度为 $n-1$ 的学生 t 分布，即

$$t=\frac{\bar{x}-\mu}{s/\sqrt{n}}\sim t(n-1)$$

可以根据学生 t 分布来估计总体均值。在给定置信水平的情况下，总体均值 μ 的估计区间为

$$\left[\bar{x}-t_{\alpha/2}\frac{s}{\sqrt{n}},\bar{x}+t_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

例 5.3 在例 5.2 的基础上，如果假设第一季度仅抽取了 20 天的快递件数，数据如图 5.4 所示。在 95％的置信水平下，求该快递公司第一季度日均快递件数的置信区间。

	A	B	C	D	E	F	G	H	I	J
1	890	873	679	982	1031	760	1130	965	1210	890
2	1502	1210	730	1130	965	977	1320	1452	899	1502

图 5.4 第一季度抽取的 20 天的快递件数

请注意，本例与例 5.2 的唯一不同点在于减少了样本数量，变成小样本的情况。

步骤如下：

(1) 打开“第 5 章参数估计.xlsx”文件，选择“例 5.3”工作表，在 A1:J2 单元格区域中输入图 5.4 中的数据。

(2) 计算样本均值。在 A4 单元格中输入“样本均值”，在 B4 单元格中输入公式“＝AVERAGE(A1:J2)”。

(3) 计算标准差。在 A5 单元格中输入“标准差”，在 B5 单元格中输入公式“＝STDEV.S(A1:J2)”

(4) 计算样本容量。在 A6 单元格中输入“样本容量”，在 B6 单元格中输入公式“＝COUNT(A1:J2)”。

(5) 在 A7 单元格中输入“置信水平”，在 B7 单元格中输入 0.95。

(6) 计算标准误差。在 A9 单元格中输入“标准误差”，在 B9 单元格中输入公式“＝B5/SQRT(B6)”。

(7) 计算 T 值。在 A10 单元格中输入“T 值”，在 B10 单元格中输入公式“＝T.INV(1－B7,B6－1)”。

(8) 计算极限误差。在 A12 单元格中输入“极限误差”，在 B12 单元格中输入公式“＝B9 * B10”。

(9) 计算置信上限与置信下限。在 A14 单元格中输入“置信上限”，在 B14 单元格中输入公式“＝B4＋B12”。在 A15 单元格中输入“置信下限”，在 B15 单元格中输入公式“＝B4－B12”。

最终结果如图 5.5 所示。

	A	B	C	D	E	F	G	H	I	J
1	890	873	679	982	1031	760	1130	965	1210	890
2	1502	1210	730	1130	965	977	1320	1452	899	1502
3										
4	样本均值	1054.85								
5	标准差	247.1816								
6	样本容量	20								
7	置信水平	0.95								
8										
9	标准误差	55.2715								
10	T值	2.093024								
11										
12	极限误差	115.6846								
13										
14	置信上限	1170.535								
15	置信下限	939.1654								

图 5.5　例 5.3 最终结果

下面使用 Excel 提供的“函数”对话框计算置信区间。

例 5.4　图 5.6 为某地区 15 户家庭年收入数据(单位：万元)。在 95%的置信水平下,求该地区家庭年收入均值的置信区间。

	A	B	C	D	E
1	20.25	15.32	14.28	17.93	25.98
2	10.28	13.27	18.74	16.23	14.38
3	12.39	11.68	10.42	11.72	16.54

图 5.6　家庭年收入数据

步骤如下：

(1) 打开“第 5 章参数估计.xlsx”文件,选择“例 5.4”工作表,在 A1:E3 单元格区域中输入图 5.6 中的数据。

(2) 计算样本均值。在 A5 单元格中输入“样本均值”,在 B5 单元格中输入公式“=AVERAGE(A1:E3)”。

(3) 计算标准差。在 A6 单元格中输入“标准差”,在 B6 单元格中输入公式“=STDEV.S(A1:E3)”

(4) 计算样本容量。选择 A7 单元格中输入“样本容量”,在 B7 单元格中输入公式“=COUNT(A1:E3)”。

(5) 在 A8 单元格中输入“置信水平”,在 B8 单元格中输入 0.95。

(6) 计算标准误差。在 A10 单元格中输入“标准误差”,在 B10 单元格中输入公式“=B5/SQRT(B6)”。

(7) 计算 T 值。在 A11 单元格中输入“T 值”。选中 B11 单元格,在功能区中单击 *fx* 按钮,打开“插入函数”对话框,如图 5.7 所示,在“或选择类别”下拉菜单中选择“全部”选项,在“选择函数”下拉菜单中找到 T.INV.2T 函数,单击“确定”按钮,打开 TINV 的“函数参数”对话框。在 Probability 文本框中输入 1−B8,在 Deg_freedom 文本框中输入 B7−1,如图 5.8 所示,单击“确定”按钮,得到 T 值。

(8) 计算极限误差。在 A13 单元格中输入“极限误差”,在 B13 单元格中输入公式“=B10 * B11”。

图 5.7 “插入函数”对话框

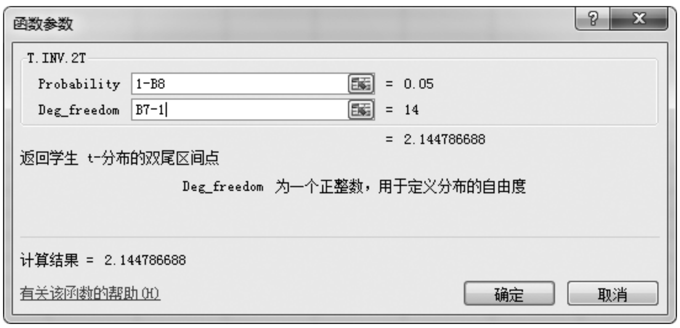

图 5.8 T.INV.2T 的“函数参数”对话框

(9) 计算置信上限与置信下限。在 A15 单元格中输入“置信上限”，在 B15 单元格中输入公式“=B5+B13”。在 A16 单元格中输入“置信下限”，在 B16 单元格中输入公式“=B5−B13”。

最终结果如图 5.9 所示。

	A	B	C	D	E
1	20.25	15.32	14.28	17.93	25.98
2	10.28	13.27	18.74	16.23	14.38
3	12.39	11.68	10.42	11.72	16.54
4					
5	样本均值	15.294			
6	标准差	4.218405			
7	样本容量	15			
8	置信水平	0.95			
9					
10	标准误差	1.089187			
11	T值	2.144787			
12					
13	极限误差	2.336075			
14					
15	置信上限	17.63007			
16	置信下限	12.95793			

图 5.9 例 5.4 最终结果

由上述两个例子可以发现，计算总体方差未知且为小样本情况下的均值置信区间的步骤与计算总体方差已知或大样本情况下的均值置信区间的步骤几乎一样。不同之处在于：在总体方差未知或小样本情况下需要使用 T.INV.2T 函数计算 T 值，而在总体方差已知或大样本情况下使用 NORM.S.INV 函数计算 Z 值。总体方差未知且小样本情况下的均值置信区间的通用计算步骤如下：

（1）利用给定的数据计算样本均值、标准差、样本容量和置信水平。

（2）根据步骤（1）中的标准差和样本容量计算标准误差，计算公式为

标准误差＝标准差/SQRT（样本容量）

（3）根据步骤（1）中的置信水平和样本容量计算 T 值，计算公式为

T 值＝T.INV.2T（1－置信水平，样本容量－1）

（4）根据步骤（2）求出的标准误差和步骤（3）求出的 T 值计算极限误差，计算公式为

极限误差＝标准误差×T 值

（5）根据步骤（1）中的样本均值和步骤（4）求出的极限误差计算置信区间，计算公式为

置信上限＝样本均值＋极限误差

置信下限＝样本均值－极限误差

5.2.3 重复抽样情况下的总体比例区间估计

总体比例分布直接来源于二项分布，但是当 n 较大时，概率计算比较复杂。根据中心极限定理，随着样本容量的增加，二项分布逐渐接近正态分布。一般来说，如果 $np \geqslant 5$ 并且 $n(1-p) \geqslant 5$，总体比例的抽样分布可以用正态分布近似。在给定置信水平的情况下置信区间为

$$\left[p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right]$$

其中，p 为样本比例，n 为样本容量。

例 5.5 某市随机抽取了当年 300 名新生儿，其中有 183 名男婴。在 95％的置信水平下，求该市男婴比例的置信区间。

步骤如下：

（1）打开“第 5 章参数估计.xlsx”文件，选择“例 5.5”工作表，在 A1:B3 单元格区域中输入图 5.10 中的数据。其中，B2 单元格中的计算公式为“＝183/300”。

	A	B
1	n	300
2	p	0.61
3	置信水平	0.95

图 5.10 男婴比例数据

（2）计算标准误差。在 A5 单元格中输入“标准误差”，在 B5 单元格中输入公式“＝SQRT(B2 * (1－B2)/B1)”。

（3）计算 Z 值。在 A6 单元格中输入“Z 值”，在 B6 单元格中输入公式“＝ABS(NORM.S.INV((1－B3)/2))”。

（4）计算极限误差。在 A8 单元格中输入“极限误差”，在 B8 单元格中输入公式“＝B5 * B6”。

(5) 计算置信上限与置信下限。在 A10 单元格中输入“置信上限”,在 B10 单元格中输入公式“＝B2＋B8”。在 A11 单元格中输入“置信下限”,在 B11 单元格中输入公式“＝B2－B8”。

最终结果如图 5.11 所示。

	A	B
1	n	300
2	p	0.61
3	置信水平	0.95
4		
5	标准误差	0.02816
6	Z值	1.959964
7		
8	极限误差	0.055193
9		
10	置信上限	0.665193
11	置信下限	0.554807

图 5.11　例 5.5 最终结果

5.3　必要样本容量计算

在实际问题中,通常需要首先确定样本容量。如果样本容量大,通常估计的精确性能有提高,但是也意味着成本的增加;如果样本容量小,成本可以有所节省,但是估计的精确性不能得到保障。故而,选取适当的样本容量,使之既能保障估计的精确性,又不至于使成本过高,就成了每一位数据统计人员必须考虑的问题。一般根据给定的极限误差计算样本容量。在本节中,分别讨论总体方差已知或大样本的情况下和重复抽样总体比例已知的情况下,必要样本的容量计算。

5.3.1　总体方差已知或大样本情况下的必要样本容量计算

假设抽样极限误差为 e,则求极限误差的公式为

$$e = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

通过上述公式可以反推出求样本容量 n 的公式:

$$n = \frac{z_{\alpha/2}^2\sigma^2}{e^2}$$

例 5.6　某公司要了解快递行业的收入情况。假设标准差为 100 元。在 95%的置信水平下,需要多少样本才能让极限误差不超过 30 元?

	A	B
1	标准差	100
2	极限误差	30
3	置信水平	0.95

图 5.12　输入数据

步骤如下:

(1) 打开“第 5 章参数估计.xlsx”文件,选择“例 5.6”工作表,在 A1:B3 单元格区域中输入图 5.12 中的数据。

(2) 计算 Z 值。在 A5 单元格中输入“Z 值”,在 B5 单元格中

输入公式"＝ABS(NORM.S.INV((1－B3)/2))"。

(3) 计算样本容量。在A6单元格中输入"样本容量"，在B6单元格中输入公式"＝B5^2＊B1^2/B2^2"。

(4) 取整。在A7单元格中输入"取整"，在B7单元格中输入公式"＝CEILING(B6，1)"。

最终结果如图5.13所示。

	A	B
1	标准差	100
2	极限误差	30
3	置信水平	0.95
4		
5	Z值	1.959964
6	样本容量	42.68288
7	取整	43

图5.13　例5.6最终结果

通过例5.6可以发现，样本容量其实就是通过极限误差的计算公式反向求出的。

可以通过修改例5.6中极限误差的值，观察样本容量的变化，例如，观察极限误差分别取50和20时样本容量的变化。

5.3.2　重复抽样总体比例已知情况下的必要样本容量计算

假设抽样极限误差为e，则求极限误差的公式为

$$e = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

根据该公式可以反推出求样本容量的公式：

$$n = \frac{p(1-p)z_{\alpha/2}^2}{e^2}$$

例5.7　某公司要调查某一产品的消费者满意度。公司确信消费者的满意度不低于60％。如果希望极限误差为2％，在置信水平为95％的情况下，需要多大的样本容量？

(1) 打开"第5章参数估计.xlsx"文件，选择"例5.7"工作表，在A1:B3单元格区域中输入图5.14中的数据。

(2) 计算Z值。在A5单元格中输入"Z值"，在B5单元格中输入公式"＝ABS(NORM.S.INV((1－B3)/2))"。

(3) 计算样本容量。在A6单元格中输入"样本容量"，在B6单元格中输入公式"＝B1＊(1－B1)＊B5^2/B2^2"。

(4) 取整。在A7单元格中输入"取整"，在B7单元格中输入公式"＝CEILING(B6，1)"。

最终结果如图5.15所示。

	A	B
1	比例	60%
2	极限误差	2%
3	置信水平	0.95

图5.14　输入数据

	A	B
1	比例	60%
2	极限误差	2%
3	置信水平	0.95
4		
5	Z值	1.959964
6	样本容量	2304.875
7	取整	2305

图5.15　例5.7最终结果

由例5.6和例5.7可以发现，两种情况下计算必要样本容量的步骤基本相同，区别在于求样本容量的公式不同。求必要样本容量的通用步骤可概括如下：