测绘科技专著出版基金资助

CEHUI KEJI ZHUANZHU CHUBAN JIJIN ZIZHU

STATISTIC THEORY AND METHOD OF SURVEYING DATA PROCESSING

陶本藻 编著

测量数据处理的统计理论和方法

测绘出版社

测绘科技专著出版基金资助

测量数据处理的统计理论和方法

Statistic Theory and Method of Surveying Data Processing

陶本藻　编著

测绘出版社

·北京·

内 容 简 介

本书融作者多年的教学成果和科研成果，从测量数据统计特性出发，叙述测量数据建模的理论及其统计处理方法。全书共分7章，其内容包括线性模型的参数估计和假设检验，介绍必要的统计理论和现今的有效处理方法；对现代测量数据处理中出现的模型误差，介绍了作者多年来的研究成果，并提出了进一步研究的方向；最后讨论了应用最为广泛的动态数据检测网的理论及其统计处理的分析方法。

本书可作为测绘工程专业本科生辅导用书和从事数据处理等相关专业人员参考书。

图书在版编目(CIP)数据

测量数据处理的统计理论和方法/陶本藻编著. —北京：测绘出版社，2007.9

ISBN 978-7-5030-1716-2

Ⅰ. 测… Ⅱ. 陶… Ⅲ. 测量—数据处理 Ⅳ. P207

中国版本图书馆 CIP 数据核字(2007)第 132144 号

责任编辑	周广森 田 力	封面设计	赵培璧
出版发行	测绘出版社		
社 址	北京西城区复外三里河路50号	邮政编码	100045
电 话	010－68512386 68531558	网 址	www.sinomaps.com
印 刷	北京通州区次渠印刷厂	经 销	新华书店
成品规格	169 mm×239 mm	印 张	15.75
字 数	300千字		
版 次	2007年9月第1版	印 次	2007年9月第1次印刷
印 数	0001－2000册	定 价	36.00元

书 号 978-7-5030-1716-2/P·457

如有印装质量问题，请与我社发行部联系

前 言

测量数据处理的对象是带有不可避免误差的观测值，19 世纪初发展起来的最小二乘法和测量平差，是基于观测误差服从高斯的误差定律所建立起来的正态分布统计理论。随着测量平差独立学科的形成，概率统计理论得到了广泛和深入地应用。20 世纪 50 年代，契巴达廖夫的著作《最小二乘法与概率论基础》(1958)是这种结合最早的一本教科书。李庆海、陶本藻在《概率统计原理和在测量中的应用》(1982)一书的前言中指出，“这是一种值得鼓励的结合，可以使最小二乘法理论更为周密，内容更为丰富”。概率统计与测量平差最为深入和全面的结合，Koch 的专著《线性模型的参数估计和假设检验》(1980)应该是最好的作品。虽然概率统计的理论和方法在测量平差中得到充分应用，由于测量学科中的平差模型远比概率统计中出现的模型要复杂和多样，测绘学者在应用的同时又对概率统计的理论进行了许多扩展和补充。

随着测绘科学与技术及其相关学科的迅速发展和生产实践高精度的需求，所研究的观测误差特性已从偶然误差扩展至系统误差和粗差，数据处理对象已从局限于静态估计扩展至随时间变化的动态估计。由此出现的新理论和新方法极大地丰富和充实了测量数据统计处理的内容。本书将概率统计和现代测量平差融为一体，将概率统计模型化，测量平差模型统计化，既充实了平差理论的完整性，又扩展了概率统计的基本理论及应用。本书重点是阐述现代测量平差的统计理论和方法，而上述的《概率统计原理和在测量中的应用》则是重点讲述数理统计理论并应用于测量平差。

本书所阐述的内容主要是本人学习和研究成果。《测量数据统计分析》(1991)是反映笔者本人当时研究成果的专著，将其中主要的并对当前仍有理论和应用价值的成果也收录于本书中。

本书的第 1 章和第 2 章是测量数据处理中常用的概率分布与数字特征理论，也是平差估计和精度评定的理论基础。第 3 章阐述按广义逆作测量平差的方法，从平差应用角度对广义逆的性质和运算作了较全面介绍，这在现今的测量平差书籍中是难以见到的。第 4 章对现有的各种平差线性模型进行了分析，从最优线性无偏估计出发给出了参数估计理论和方法。第 5 章线性模型的统计假设检验，将数理统计中的回归分析和方差分析模型化，导出了统计检验公式，还给出了作者提出的线性假设法的扩充。第 6 章和第 7 章是作者主要研究方向。第 6 章模型误差及其统计分析，重点阐述作者提出的模型误差识别、估计和补偿的理论和方法。第 7 章变形模型统计分析，由于本人已出版教材《自由网平差与变形分析》(1984，2001)，为不与该书重复，仅写出该书未列入的成果。

由于本书所述内容，都是国内外测绘界感兴趣的热门前沿课题，发表的论文、文献特别多，在编写本书时作过大量阅读，并深受启发。但本书在参考文献中没有一一列出，而仅按本书所引用的文献，以先后为序列出其中一部分，也对此作一说明。

本书的出版得到国家测绘局测绘科技专著出版基金资助，测绘出版社为本书编辑出版做了大量工作，作者深表感谢。

作者

2007 年 4 月

目　　录

Contents

第1章　绪　论

测量数据处理的对象是带有不可避免误差的观测量，观测误差是随机变量，任何随机变量都有一定的概率分布。本章在阐述高斯误差分布定律基础上，给出概率分布重要特征值：数学期望、方差和协方差的定义和运算规则，说明在测量数据处理中含义及应用。最后简述数理统计在测量数据处理中应用的内容，并指出在应用过程中测绘界也拓展了数理统计理论和方法。

§1.1　观测误差及其概率分布

一、观测误差为随机变量

一种随机实验的结果，当用数字表达出来时，则称为随机变量，一般它以不同的概率取不同的数值。观测是一种随机实验，观测结果事先不能预料，故观测量是一个随机变量，这是由于不可避免地观测误差之故。由于观测误差这一随机变量其大小取值是不可列的，它在可能取值范围内充满区间，或者说在此区间内有无穷多个连续点，所以观测误差（或观测量）为连续型随机变量。观测误差定义为真值 $\overline{X}$ 与观测值 L 之差，即 $\Delta=\overline{X}-L$。设其概率密度函数为 $f(\Delta)$，则有

$$\int_{-\infty}^{\infty} f(\Delta)\mathrm{d}\Delta = 1$$

概率分布函数为

$$F(x) = P(\Delta \leqslant x) = \int_{-\infty}^{x} f(\Delta)\mathrm{d}\Delta$$

由此可得误差 Δ 在任意区间$[a\quad b]$取值的概率为

$$P(a < \Delta \leqslant b) = F(b) - F(a) = \int_{a}^{b} f(\Delta)\mathrm{d}\Delta$$

二、高斯误差分布定律

为了推导偶然误差分布密度函数 $f(\Delta)$ 的具体形式，高斯首先给出 $f(\Delta)$ 应具有的性质。

偶然误差第一特性：在一定观测条件下，偶然误差的绝对值不会超过一定的限值。设此限值为 B，则误差出现在$[-B\quad B]$之内是必然事件，即其概率 $P(-B<\Delta \leqslant B) = 1$，故有

$$P(-B < \Delta \leqslant B) = \int_{-B}^{B} f(\Delta)\mathrm{d}\Delta = 1$$

这是 $f(\Delta)$ 所应具有的一个性质。

误差第二特性：绝对值较小的误差比绝对值较大的误差出现的概率较大，即当 $|\Delta_1| > |\Delta_2|$ 时，则

$$P(\Delta_1) = f(\Delta_1)\mathrm{d}\Delta < P(\Delta_2) = f(\Delta_2)\mathrm{d}\Delta$$

或

$$f(\Delta_1) < f(\Delta_2)$$

就是说，$f(\Delta)$ 在负半开区间 $[-\infty, 0]$ 中是单调增的，而在正半开区间 $[0, \infty]$ 是单调减的。

误差第三特性：绝对值相等的正误差与负误差出现的概率相等。由此知

$$f(\Delta)\mathrm{d}\Delta = f(-\Delta)\mathrm{d}\Delta$$

或

$$f(\Delta) = f(-\Delta)$$

就是说，$f(\Delta)$ 应是偶函数。

仅满足以上三个性质，还不足以确定惟一的 $f(\Delta)$ 形式。高斯提出再增加一个条件。设对某量 $\overline{X}$ 作 n 次独立同精度观测，得 $L_1, L_2, \cdots, L_n$，其相应的误差为 $\Delta_i = \overline{X} - L_i$。真值 $\overline{X}$ 未知，取观测值的算术平均值 $X = \frac{1}{n}\sum_{i=1}^{n} L_i$ 为 $\overline{X}$ 的最或然估值。增加此条件就可导出 $f(\Delta)$ 的具体形式了。

所谓最或然值 X，是指由下列最大或然法准则

$$G = f(X - L_1) f(X - L_2) \cdots f(X - L_n) = 最大$$

求出的 $\overline{X}$ 的估值，上述准则也可写成

$$\ln G = \ln f(X - L_1) + \ln f(X - L_2) + \cdots + \ln f(X - L_n) = 最大 \tag{1-1}$$

下面介绍高斯推导 $f(\Delta)$ 公式的方法。

能使式(1-1)成立的 X 值，必须使其一阶导数为零，即

$$\frac{f'(X - L_1)}{f(X - L_1)} + \frac{f'(X - L_2)}{f(X - L_2)} + \cdots + \frac{f'(X - L_n)}{f(X - L_n)} = 0$$

将式中的每一项记为 $\varphi(X - L_i)$，$i = 1, 2, \cdots, n$，则有

$$\varphi(X - L_1) + \varphi(X - L_2) + \cdots + \varphi(X - L_n) = 0 \tag{1-2}$$

这就是符合最大或然法的 X 值应满足的方程。

以算术平均值作为最或然值，下式必然成立，即

$$(X - L_1) + (X - L_2) + \cdots + (X - L_n) = 0 \tag{1-3}$$

此即 $nX - \sum_{i=1}^{n} L_i = 0$，式(1-2)和式(1-3)应同时成立。令 $V_i = X - L_i$，式(1-2)、式(1-3)分别写成

$$\varphi(V_1)+\varphi(V_2)+\cdots+\varphi(V_n)=0$$
$$V_1+V_2+\cdots+V_n=0$$

将 $\varphi(V_i)$ 表示为幂级数

$$\varphi(V_i)=K_0+K_1V_i+K_2V_i^2+K_3V_i^3+\cdots \quad (i=1,2,\cdots,n)$$

式中 $K_j(j=0,1,2,\cdots)$ 为常数，取上列 n 式之和，并顾及式(1-2)得

$$0=nK_0+K_1\sum_{i=1}^{n}V_i+K_2\sum_{i=1}^{n}V_i^2+K_3\sum_{i=1}^{n}V_i^3+\cdots$$

由式(1-3)知 $\sum_{i=1}^{n}V_i=0$，因 $n\neq 0$，$\sum_{i=1}^{n}V_i^2$，$\sum_{i=1}^{n}V_i^3$ 不恒等于零。若上式永远成立，则式中的 $K_0, K_2, K_3, \cdots$ 应必为零。又因 $\varphi(V_i)$ 不恒为零，故 K_1 不应为零，由此知

$$\varphi(V_i)=K_1V_i \qquad (i=1,2,\cdots,n)$$

则式(1-2)与式(1-3)的对应项必然成比例，即有

$$\frac{\varphi(V_1)}{V_1}=\frac{\varphi(V_2)}{V_2}=\cdots=\frac{\varphi(V_n)}{V_n}=K$$

由于不管 V 为何值，上式均可成立，故一般的可写成

$$\frac{\varphi(\Delta)}{\Delta}=K$$

这就是以算术平均值作为 $\overline{X}$ 的最或然值时 $f(\Delta)$ 所应满足的关系式。因

$$\varphi(\Delta)=\frac{f'(\Delta)}{f(\Delta)}=\frac{df(\Delta)}{f(\Delta)d\Delta}=K\Delta$$

对上式最后等式两边求积分

$$\ln f(\Delta)=\frac{1}{2}K\Delta^2+C$$

C 为积分常数，上式写成指数形式为

$$f(\Delta)=\mathrm{e}^{(\frac{1}{2}K\Delta^2+C)}=\mathrm{e}^c\mathrm{e}^{\frac{1}{2}K\Delta^2}=A\mathrm{e}^{\frac{1}{2}K\Delta^2}$$

式中用符号 A 代常数 e^c。

由误差第二、第三特性知，$f(\Delta)$ 为偶函数，且其值随误差 $|\Delta|$ 的增大而减小，因而上式中的指数应为负数，于是令

$$\frac{K}{2}=-h^2$$

则有

$$f(\Delta)=A\mathrm{e}^{-h^2\Delta^2}$$

由误差第一特性，$f(\Delta)$ 还应满足下式：

$$\int_{-B}^{B}A\mathrm{e}^{-h^2\Delta^2}\,\mathrm{d}\Delta=1$$

由此确定积分常数 A 还有困难，因在不同的观测条件有不同的 B 值，即使对特定的

观测条件而言，B 也难于精确给定。考虑到函数在 $|\Delta|$ 较大时，$f(\Delta)$ 实际上趋近于零，因此可将积分上、下限扩展到 $-\infty$ 至 ∞，使有可能利用普阿桑积分确定 A 值。这样上式为

$$\int_{-\infty}^{\infty} A\mathrm{e}^{-h^2\Delta^2}\,\mathrm{d}\Delta = 1$$

现令 $t = h\Delta$，故 $\mathrm{d}t = h\mathrm{d}\Delta$，由此上式为

$$\int_{-\infty}^{\infty} A\mathrm{e}^{-h^2\Delta^2}\,\mathrm{d}\Delta = \frac{A}{h}\int_{-\infty}^{\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t = 1$$

根据普阿桑积分

$$\int_{-\infty}^{\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t = \sqrt{\pi}$$

可得

$$A = \frac{h}{\sqrt{\pi}}$$

由此即得 Δ 的分布密度函数为

$$f(\Delta) = \frac{h}{\sqrt{\pi}}\mathrm{e}^{-h^2\Delta^2} \tag{1-4}$$

因为概率没有负值，$f(\Delta)$ 是非负的，故上式中的 $h>0$。式(1-4) 即为高斯误差分布定律。

h 为精度指标，h 与标准差 σ 的关系为

$$h = 1/\sqrt{2}\sigma$$

式(1-4) 的另一形式为

$$f(\Delta) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\Delta^2/2\sigma^2} \tag{1-5}$$

下面给出证明。

将式(1-4) 代入最大或然准则得

$$G = \left(\frac{h}{\sqrt{\pi}}\right)^n \mathrm{e}^{-h^2\sum\limits_{i=1}^{n}\Delta_i^2}$$

令 $\mathrm{d}G/\mathrm{d}h = 0$ 得

$$\left(\frac{\hat{h}}{\sqrt{\pi}}\right)^n \mathrm{e}^{-\hat{h}^2\sum\limits_{i=1}^{n}\Delta_i^2}\left(-2\hat{h}\sum_{i=1}^{n}\Delta_i^2\right) + \mathrm{e}^{-\hat{h}^2\sum\limits_{i=1}^{n}\Delta_i^2} \cdot \frac{n}{(\sqrt{\pi})^n}\hat{h}^{n-1} = 0$$

$$\left(\frac{\hat{h}}{\sqrt{\pi}}\right)^n \mathrm{e}^{-\hat{h}^2\sum\limits_{i=1}^{n}\Delta_i^2}\left(-2\hat{h}\sum_{i=1}^{n}\Delta_i^2 + \frac{n}{\hat{h}}\right) = 0$$

或

$$\left(-2\hat{h}\sum_{i=1}^{n}\Delta_i^2 + \frac{n}{\hat{h}}\right) = 0$$

于是 h 的最或然值为

$$\hat{h}^2 = \frac{n}{2\sum_{i=1}^{n}\Delta_i^2} = \frac{1}{2\hat{\sigma}^2}$$

式中 $\hat{\sigma}^2 = \sum_{i=1}^{n}\Delta_i^2/n$。当 n 趋于无穷时，$\hat{\sigma}^2$ 趋于理论值 σ^2，$\hat{h}^2$ 也趋于 h^2，故有 $h = 1/\sqrt{2}\sigma$。

由式(1-5)知，当 $\Delta = 0$ 时，$f(\Delta) = 0$。以 Δ 为横轴，$f(\Delta)$ 为纵轴，坐标原点为零，$f(\Delta)$ 对称于纵轴。$f(\Delta)$ 有两个参数。一是对称轴的横坐标，是该分布的数学期望，此处为零，记为 $E(\Delta) = 0$。一是方差 σ^2，$f(\Delta)$ 是正态分布的密度函数，故观测误差 Δ 服从正态分布，简记为 $\Delta \sim N(0,\sigma^2)$。

§1.2　数学期望

一、定义

一维分布中，以随机变量 X 所取数值的相应概率为权，求得加权平均值，称为随机变量 X 的数学期望或称该分布的均值，记为 $E(X)$ 或 ξ。

对于离散型分布列 $P(X = x_i) = p_i$，$\sum_{i=1}^{n} p_i = 1$，其数学期望为

$$E(X) = \frac{p_1x_1 + p_2x_2 + \cdots + p_nx_n}{p_1 + p_2 + \cdots + p_n} = \sum_{i=1}^{n} p_ix_i \Big/ \sum_{i=1}^{n} p_i$$

对于连续型分布，则为

$$E(X) = \int_{-\infty}^{\infty} xf(x)\mathrm{d}x \tag{1-6}$$

因为上述求数学期望的方法与力学中求质量重心坐标的方法一致，所以数学期望也可以看作分布重心的横坐标。

为了描述一个分布，除了密度函数和分布函数外，还可用分布的特征值来描述。数学期望是主要特征值之一。它描述分布沿横轴的位置，作为位置特征。

二、性质

1. 设 k 为常数，则

$$E(k) = k$$

2. 设 k 为常数，则

$$E(kX) = kE(X)$$

以上两个性质，顾及$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$，可直接由定义式(1-6)得出。

3. 无论各变量独立与否，其和的数学期望等于每一变量数学期望之和，即

$$E(X+Y) = E(X) + E(Y)$$

证明如下：

$$E(X+Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f(x,y)\mathrm{d}x\mathrm{d}y = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf(x,y)\mathrm{d}x\mathrm{d}y + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf(x,y)\mathrm{d}x\mathrm{d}y = \int_{-\infty}^{\infty}xf_1(x)\mathrm{d}x + \int_{-\infty}^{\infty}yf_2(y)\mathrm{d}y = E(X) + E(Y)$$

式中 $f_1(x)$，$f_2(y)$ 为边际密度函数，一般

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

4. 如各变量互独立，则其乘积的数学期望等于各变量数学期望之乘积，即

$$E(X_1X_2\cdots X_n) = E(X_1)E(X_2)\cdots E(X_n)$$

这是因为

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyf(x,y)\mathrm{d}x\mathrm{d}y = \int_{-\infty}^{\infty}xf_1(x)\mathrm{d}x\int_{-\infty}^{\infty}yf_2(y)\mathrm{d}y = E(X)E(Y)$$

三、真值与数学期望

设观测量 X 的真值为 $\overline{X}$，观测偶然误差为 Δ，则有

$$X = \overline{X} + \Delta$$

两边取数学期望，按性质 1，并顾及 $E(\Delta) = 0$，则有

$$E(X) = \overline{X}$$

因此，测量中的真值理论上可用观测量这一随机变量的数学期望来定义。

设 $\overline{X}$ 的观测值为 $x_1, x_2, \cdots, x_n$，相应的真误差为 $\Delta_1, \Delta_2, \cdots, \Delta_n$，则有

$$\begin{aligned} x_1 &= \overline{X} + \Delta_1 \\ x_2 &= \overline{X} + \Delta_2 \\ &\vdots \\ x_n &= \overline{X} + \Delta_n \end{aligned}$$

取和得

$$\sum_{i=1}^{n} x_i = n\overline{X} + \sum_{i=1}^{n}\Delta_i$$

$$\hat{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{X} + \frac{\sum_{i=1}^{n}\Delta_i}{n}$$