附件

# 环境与健康横断面调查数据统计分析技术指南

环境保护部

二〇一七年十一月

# 目　次

# 前　言

为贯彻《中华人民共和国环境保护法》，科学分析环境污染与人群健康之间的关系，指导和规范环境与健康横断面调查数据统计分析工作，制定本指南。

本指南规定了环境与健康横断面调查数据统计分析的工作流程、分析内容、分析方法和技术要求。

本指南由环境保护部科技标准司组织制订。

本指南为首次发布。

本指南主要起草单位：北京师范大学、环境保护部华南环境科学研究所、环境保护部环境与经济政策研究中心、中国医学科学院基础医学研究所（北京协和医学院基础学院）。

本指南由环境保护部 2017 年 11 月 28 日批准。

本指南自发布之日起实施。

本指南由环境保护部解释。

# 环境与健康横断面调查数据统计分析技术指南

## 1 适用范围

本指南规定了环境与健康横断面调查数据统计分析的工作流程、分析内容、分析方法和技术要求。

本指南适用于对环境与健康横断面调查所获得的数据进行描述性统计分析和推断性统计分析。

## 2 规范性引用文件

本指南引用了下列文件中的条款。凡是不注明日期的引用文件，其有效版本（包括修改单）适用于本指南。

GB/T 4882 数据的统计处理和解释 正态性检验

GB/T 4883 数据的统计处理和解释 正态样本离群值的判断和处理

GB/T 8170 数值修约规则与极限数值的表示和判定

## 3 术语和定义

3.1

**统计描述 statistical description**

利用样本统计量对总体参数进行估计，通常包括对集中趋势和变异程度的描述。

3.2

**统计推断 statistical inference**

由样本数据的特征推断总体特征。

3.3

**总体 population**

根据研究目的确定的所有同质观察单位的集合。

3.4

**样本 sample**

根据研究目的从研究总体中抽取的部分观察单位。

3.5

**参数 parameter**

反映总体特征的统计指标。

3.6

**统计量 statistic**

反映样本特征的统计指标。

3.7

**计量资料 measurement data**

按观察单位观察值的大小表示其结果的资料类型。

3.8

**计数资料 count data**

又称分类资料，指按观察单位不同属性进行分组计数的资料类型。

3.9

**等级资料 ordinal data**

将观察单位按某种属性的不同程度或次序分成等级后分组计数的结果，具有半定量性质。

3.10

**正态性检验 normality test**

判断样本所在总体是否服从正态分布的一种假设检验方法。

3.11

**平均数 average**

描述一组计量资料集中趋势的统计指标，常用以说明该组数据的平均水平。表示资料集中趋势的几个比较重要的指标有算术均数、几何均数、中位数。

3.12

**算术均数 arithmetic mean**

所有观察值之和与观察值个数的商。

3.13

**几何均数 geometric mean**

所有观察值对数的算术均数的反对数。

3.14

**中位数 median**

将所有观察值从小到大排列后处于中间位置的值。如果观察值的个数为偶数，中位数则为中间两个数的算术均数。

3.15

**方差 variance**

描述一个变量所有观察值与总体均数离散程度的指标，是各个观察值与算术均数之差的平方和与观察值的个数 $n$（计算总体方差）或 $n$-1（计算样本方差）的商。

3.16

**标准差 standard deviation**

描述一个变量所有观察值与总体均数离散程度的指标，是方差的算术平方根。

3.17

**变异系数 coefficient of variation**

度量相对离散程度的指标，是一组数据的标准差与平均数的比值。

3.18

**分位数 quantile**

将样本按其数值从小到大排列，并将样本总量分为若干等份时，每一个界点位置上的样本值。

3.19

**百分位数 percentile**

将样本按其数值从小到大排列，并将样本总量分为 100 等份时，每一个界点位置上的样本值。

3.20

**四分位数间距 interquartile range，IQR**

第三四分位数（第 75 百分位数，$P_{75}$）与第一四分位数（第 25 百分位数，$P_{25}$）之差。

3.21

**相对数 relative number**

两个相关联指标的数值之比，表示二者的相对大小。根据相对数的分子与分母所选用指标性质的不同，相对数又分为率、构成比和相对比。

3.22

**率 rate**

在某一时点/时段内某现象或事件发生的频率或强度，用在某时点/时段内某现象或事件发生数与在该时点/时段内可能发生该现象或事件的观察单位总数之比表示。

3.23

**构成比 proportion**

某事物内部各个组成部分所占的比重，是各个组成部分的观察单位数与全部观察单位数的比值。

3.24

**相对比 relative ratio**

两个相关指标 A 与 B 的值之商，用以描述两者的对比水平，说明 A 与 B 的倍数关系。

3.25

**假设检验 hypothesis testing**

基于数理统计学原理，根据一定的假设条件由样本推断总体的一种方法。常用的假设检验方法有 $Z$ 检验、$t$ 检验、方差分析、$\chi^2$ 检验和秩和检验等。

3.26

**方差齐性检验 homogeneity test of variances**

判断各样本所代表的总体的方差是否相等的方法。

3.27

***t* 检验 *t*-test**

样本量 $n<100$ 时，对总体方差未知的样本均数与总体均数间或两个样本均数间进行差异检验的方法，其基本前提为样本随机地取自正态分布的总体，且两样本均数比较时两样本所对应的两总体方差相等。

3.28

***t'*检验 *t'*-test**

样本量 $n<100$ 时，对总体方差未知的样本均数与总体均数间或两个样本均数间进行差异检验的方法，其基本前提为样本随机地取自正态分布的总体，且两样本均数比较时两样本所对应的两总体方差不相等。

3.29

***Z* 检验 *Z* test**

又称 $u$ 检验，指样本量 $n\geq100$ 时，对样本均数与总体均数间或两样本均数间进行差异检验的方法。

3.30

**方差分析 analysis of variance**

又称 $F$ 检验，用于两个及两个以上样本均数间差异的显著性检验。其基本思路是将样本观察值的总变异分解为组间变异和组内变异后，通过检验二者的差异显著性来分析影响总变异的各个因素的效应，以判断各组样本间均数的差异是否存在显著性。

3.31

**$\chi^2$ 检验 chi-square test**

用于两个及多个总体率或总体构成比的比较、两分类变量间关联性分析、百分率线性趋势检验及频数分布的拟合优度检验的方法。

3.32

**连续性校正 $\chi^2$ 检验** continuity-adjusted chi-square test

采用 $\chi^2$ 检验对两个样本率进行差异比较时，如果总例数 $n \geqslant 40$ 但某个格子的理论频数 $1 \leqslant E < 5$，对 $\chi^2$ 统计量分布的连续性进行校正后的 $\chi^2$ 检验方法。

3.33

**Fisher 确切概率检验** Fisher's exact probability test

一种直接计算概率的假设检验方法。$\chi^2$ 检验时，如果总例数 $n < 40$ 或某个格子的理论频数 $E < 1$，行列表资料 $\chi^2$ 检验结果可能会产生偏性，此时常使用 Fisher 确切概率法直接计算概率。

3.34

**Wilcoxon 秩和检验** Wilcoxon rank sum test

又称两样本秩和检验，对两组不满足 $t$ 检验条件的完全随机设计的计量资料或等级资料，检验其分别代表的总体均数有无差异的方法。

3.35

**Kruskal-Wallis H 检验** Kruskal-Wallis H test

利用多样本的秩和来检验多个独立样本是否来自同分布总体的非参数检验方法。

3.36

**Pearson 相关系数** Pearson correlation coefficient

描述服从双变量正态分布的两个变量之间线性相关性的指标，以 $r$ 表示。

3.37

**Spearman 秩相关系数** Spearman's rank correlation coefficient

描述不服从双变量正态分布、总体分布未知或原始数据用等级表示的两个变量之间的相关性的指标，以 $r_s$ 表示。

3.38

**线性回归** linear regression

用直线方程来描述因变量与一个或多个自变量之间数量依存关系的一种方法。

3.39

**Logistic 回归** Logistic regression

因变量为二分类或多分类时，研究其与一个或多个自变量之间关联性的一种广义线性回归分析方法。

3.40

**Poisson 分布** Poisson distribution

属于离散型分布，是用以描述单位时间、空间、面积等的罕见事件发生次数的概率分布。

3.41

**Poisson 回归** Poisson regression

因变量为计数数据且服从 Poisson 分布时的一种广义线性回归分析方法。

3.42

**空间自相关** spatial autocorrelation

是样本数据的一种空间结构关系，反映样本数据之间的空间依赖程度。

3.43

**等标污染负荷** equal-standard waste load

将单位时间内污染源排放到环境的废水、废气中的污染物分别“稀释”或“浓缩”到排放标准规定的最高允许含量相当的排放量，即等标排放量。

3.44

**暴露量 exposure dose**

指人体经呼吸道、消化道和皮肤等途径接触环境污染物的量。

3.45

**暴露参数 exposure factors**

指用来描述人体暴露环境污染物的特征和行为的参数，包括身体特征参数、摄入量参数、时间-活动模式参数及其他参数。

3.46

**患病率 prevalence rate**

也称现患率或流行率，指一定期间、一定人群中某病新旧病例所占的比例。按观察时间的不同分为时点患病率（一般不超过一个月）和期间患病率（一个月以上）。

3.47

**发病率 incidence rate**

一定期间、一定人群中某病新发生的病例所占的比例。

3.48

**罹患率 attack rate**

某一局限范围、短时间（以日、周、旬、月为单位）内的发病率。

3.49

**死亡率 mortality rate**

某人群在一定期间（一般指一年）内发生死亡的频率，通常用千分率或十万分率表示。

3.50

**总率 overall rate**

不分人群及地区特征、不分疾病种类的疾病率。

3.51

**专率 specific rate**

按照观察对象的不同特征，如人群的年龄、性别、职业、民族、种族、婚姻状态等，以及疾病种类分别计算的疾病率。

3.52

**粗率 crude rate**

不分观察人群内部构成，直接计算的某事件/现象发生的频率或强度。

3.53

**调整率 adjusted rate**

又称标化率，为比较不同时期/不同地区观察人群间某事件/现象发生的频率或强度，采用统一的标准以消除待比较人群间因构成不同可能带来的影响而进行标准化调整后的率。

3.54

**横断面调查 cross-sectional study**

在特定时点或时期，对污染源、环境暴露水平和相应暴露人群的健康影响同时进行的调查。

## 4 分析流程

环境与健康横断面调查数据统计分析应包括统计分析方案制定、数据预处理、数据质量评价、统计描述与统计推断、结果表达六个步骤（图 1）。

## 5　制定统计分析方案

统计分析方案应包括分析目的、数据预处理及数据质量评价方法、分析内容、分析指标、分析方法及分析结果表达方法等主要内容。

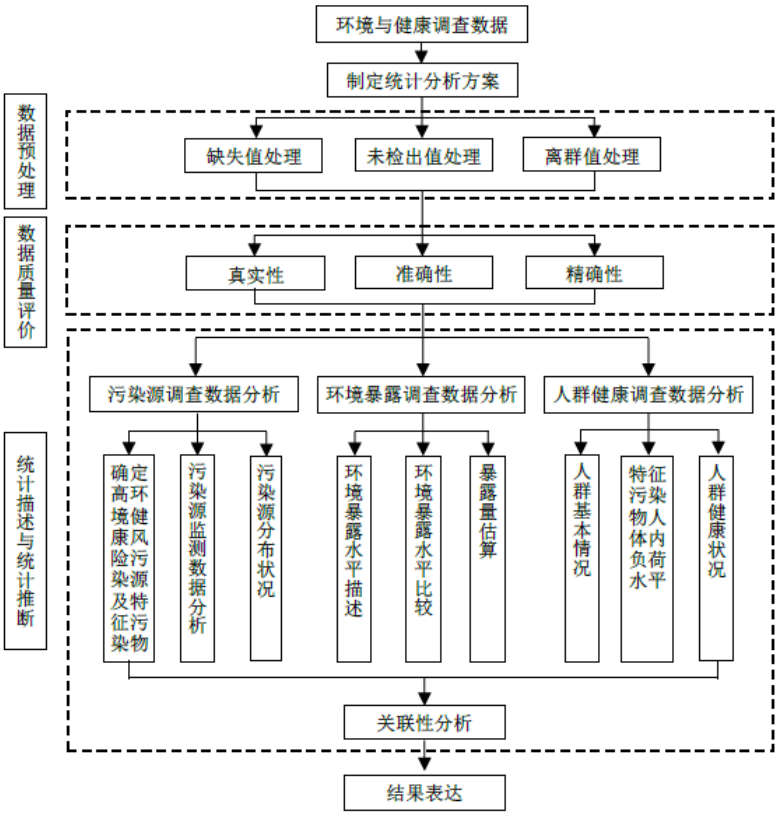


图 1 分析流程

## 6　数据预处理及数据质量评价

### 6.1　数据预处理

#### 6.1.1　缺失值处理

分析缺失值产生机制，对于可通过核实其他资料、重测、补充调查等方式弥补的，应对缺失值进行填补。如数据无法获得，可根据需要采用均值插补、回归插补、极大似然估计等统计方法对缺失值进行插补。必要时应对缺失值对分析结论的影响进行敏感性分析。

#### 6.1.2　未检出值处理

未检出值可用方法检出限的二分之一替代。

#### 6.1.3　离群值处理

应基于专业判断对离群值进行判断和核实，如为数据错误导致出现离群值的，应对数据进行修订；如数据无误但无法通过专业理论进行解释的，宜剔除离群值后进行统计分析。当样本数据服从正态分布

时，应依据 GB/T 4883 剔除离群值；当样本数据不服从正态分布时，应制作样本数据箱线图（图 2）。样本观测值记为 $X$，箱线图中满足$X<P_{25}$-1.5$IQR$ 或 $X>P_{75}$+1.5$IQR$的样本观测值为离群值，应根据实际情况予以剔除。

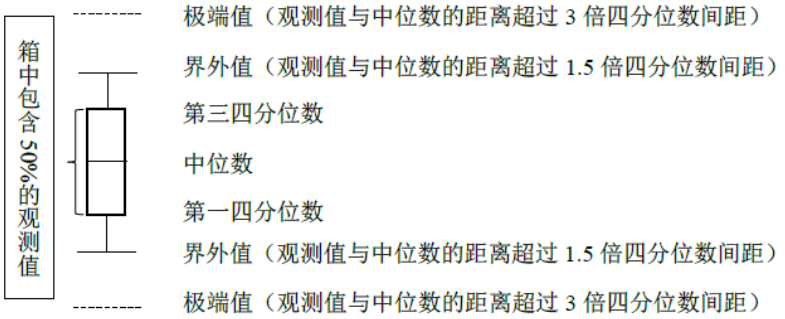


图 2 样本数据箱线图

## 6.2 数据质量评价

### 6.2.1 真实性

应采用数据的可溯源率对污染源现场调查数据、环境暴露调查数据、人群健康调查数据和实验室检测数据的真实性进行评价。数据的可溯源率为抽查的数据中与原始数据一致的记录数占抽查数据总记录数的比例。

### 6.2.2 准确性

应对采样记录数据、问卷调查数据、体格检查数据和实验室检测数据的准确性进行评价。

a）采样记录数据、问卷调查数据和体格检查数据的准确性应采用正确率进行评价，采样记录数据、问卷调查数据和体格检查数据的正确率为上报数据中无三类数据错误（缺失值、非法值、逻辑错误）中任何一项的记录数占上报的总记录数的比例。

b）实验室检测数据的准确性应采用实验室质量控制结果（如空白样品检出情况、有证标准物质检测结果、加标回收率、标准曲线相关系数等）进行评价。

### 6.2.3 精确性

应采用平行样品的相对标准偏差评价实验室检测数据的精确性。

## 7 统计分析方法选取原则

### 7.1 统计描述

#### 7.1.1 计量资料

根据数据分布类型及样本量选择统计描述指标。

a）数据服从正态分布，应采用算术均数和标准差描述其平均水平和变异程度。当样本量较大（如 $n$≥200）时，宜同时描述其最小值、$P_{25}$、中位数、$P_{75}$、最大值。

b）数据呈偏态分布但经对数转换后呈正态分布，或数据呈偏态分布但数据之间成级数关系，应采用几何均数和几何标准差描述其平均水平和变异程度。当样本量较大（如 $n$≥200）时，宜同时描述其

最小值、$P_{25}$、中位数、$P_{75}$、最大值。

c）数据呈偏态分布且不能通过数据变换转化为正态分布，应采用中位数描述其平均水平，采用四分位数间距描述其变异程度。当样本量较大（如 $n \geq 200$）时，宜同时描述其最小值、$P_{25}$、$P_{75}$、最大值。

d）当比较两组或两组以上数据变异程度时，可采用变异系数进行描述。

e）针对一些分析测试指标检出率较低（如环境介质中的有机物、人体内暴露指标）的情况，进行调查区之间比较时作如下规定：对于检出率为 100%的指标，应根据所有样品检测结果的数据分布类型选择平均数（算术均数/几何均数/中位数）描述平均水平；对于检出率高于 50%的指标，应采用所有样品检测数据的中位数描述平均水平；对于检出率低于 50%的指标，应采用检出样品的检测数据的中位数描述平均水平。

### 7.1.2 计数资料

对于计数资料的描述，可首先编制频数表，在此基础上应采用相对数（率、构成比或相对比）进行统计描述。当样本量较小（如 $n < 50$）时，应直接用分数表示。

## 7.2 统计推断

### 7.2.1 样本统计量比较

比较两个或几个样本的统计量所代表的总体参数之间的差异是否存在统计学意义时采用假设检验方法。

a）单变量计量资料的均数比较。单变量计量资料的均数比较方法应主要根据样本量、数据分布类型以及方差是否齐性等特征进行选取（图 3）。如果数据不服从正态分布或不满足方差齐性，可直接进行 *t'*检验或 Wilcoxon 秩和检验或 Kruskal Wallis *H* 检验，也可进行数据变换。如数据变换后仍不服从正态分布或不满足方差齐性，则进行 *t'*检验或 Wilcoxon 秩和检验或 Kruskal Wallis *H* 检验。

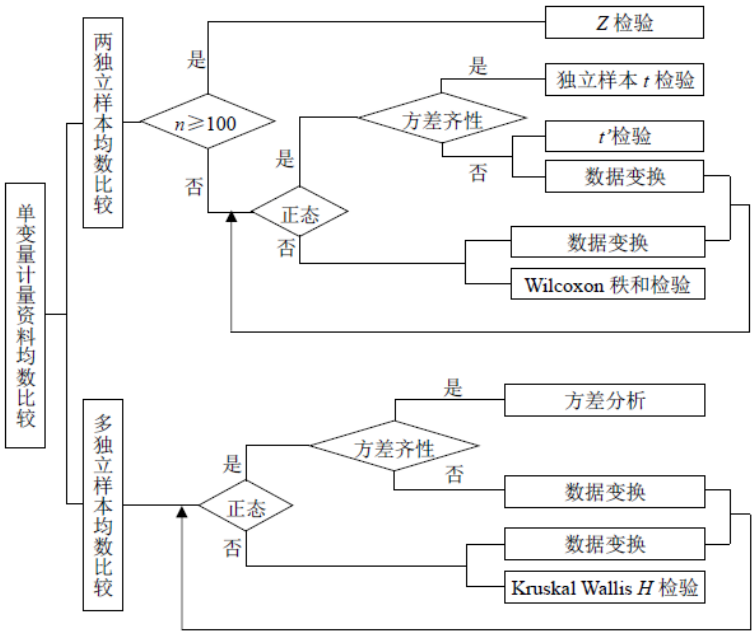


图 3 单变量计量资料均数比较方法选取程序

b）两个样本率的比较。列出四格表，当每个格子的理论频数 $E \geq 5$，且总样本例数 $n \geq 40$ 时，应采用 $\chi^2$ 检验；当总样本例数 $n \geq 40$，但其中有一个格子的理论频数 $1 \leq E < 5$ 时，应采用连续性校正 $\chi^2$ 检

验；当任何一个格子的理论频数 $E<1$，或总例数 $n<40$，或检验所得 $P$ 值接近于检验水准$\alpha$时，应采用 Fisher 确切概率检验。

c）两个样本构成比、多个样本率或构成比等相对数的比较。列出行列表，当各格子的理论频数 $E>1$，且 $E<5$ 的格子数不多于格子总数的 1/5 时，应采用 $\chi^2$ 检验；否则应采取增加观察例数等措施或进行 Fisher 确切概率检验。

#### 7.2.2 相关分析

a）对符合双变量正态分布的两变量做散点图，若呈直线或近似直线的相关关系，采用 Pearson 相关系数；

b）对不符合双变量正态分布或为等级资料的两变量的相关分析，采用 Spearman 秩相关系数。

#### 7.2.3 回归分析

a）如果因变量是连续型变量且服从正态分布、只有一个自变量、自变量和因变量相互独立、两者的散点图呈现直线或近似直线相关关系，采用一元线性回归；

b）如果因变量是连续型变量且服从正态分布、自变量超过一个、各个变量之间相互独立、因变量和各自变量之间呈现直线或近似直线相关关系，采用多元线性回归；

c）如果因变量服从或近似服从对数正态分布，先对因变量进行对数转换再进行线性回归。如果因变量服从 Poisson 分布，采用 Poisson 回归；

d）如果因变量是二分类变量或多分类变量，采用二项 Logistic 回归或多项 Logistic 回归。如果回归模型预测值与因变量值的残差存在空间自相关（莫兰指数$>0$ 且 $P<\alpha$），采用包含空间随机效应因子的回归模型；

e）回归分析中，自变量和因变量的选取应有理论依据，注意生物学可能性，不能把毫无关联的各种环境现象与健康效应进行回归。相关和回归分析的样本量过少可能导致模型估计结果的稳定性差。应结合当地实际情况和数据质量评价结果慎重地作出统计结论。

## 8 污染源调查数据分析

### 8.1 统计分析目的

a）明确污染区的高环境健康风险污染源和特征污染物，确定对照区没有污染区所关注的特征污染物的排放源；

b）通过核算排放量、分析污染物超标情况、与对照区对比等方式来明确污染源对环境质量的影响；

c）描述污染源与健康调查人群之间的位置关系。

### 8.2 数据特征

污染源调查获取的污染物排放量、排放浓度等现场监测数据以计量资料为主，其中排放浓度等现场监测数据多呈对数正态分布。

### 8.3 统计内容及方法

#### 8.3.1 确定高环境健康风险污染源及特征污染物

##### 8.3.1.1 分析内容

结合污染物毒性筛选污染区内高环境健康风险的污染源及特征污染物。

8.3.1.2 分析方法

a）等标污染负荷法。根据污染源类型、所属行业及污染物种类，选取污染物排放标准，计算污染源及污染物的等标污染负荷、等标污染负荷比（公式 1-6）。

1）等标污染负荷

$$P_{ij}=\frac{G_{ij}}{S_{ij}} \quad (1)$$

$$P_n=\sum_{i=1}^{n}P_{ij} \quad (2)$$

$$P_m=\sum_{j=1}^{m}P_{ij} \quad (3)$$

$$P=\sum_{i=1}^{n}\sum_{j=1}^{m}P_{ij} \text{ 或 } P=\sum_{j=1}^{m}\sum_{i=1}^{n}P_{ij} \quad (4)$$

式中：$P_{ij}$——某污染源（$j$）中某种污染物（$i$）的等标污染负荷；

$G_{ij}$——某污染源（$j$）中某种污染物（$i$）的年排放量；

$S_{ij}$——某污染物（$i$）的评价标准，取排放标准，不同污染源间进行比较时应选取相同的排放标准；

$P_n$——某污染源内各污染物的等标污染负荷之和；

$P_m$——污染区各污染源内某种污染物的等标污染负荷之和；

$P$——污染区内所有污染源的等标污染负荷之和。

2）等标污染负荷比

将污染物等标污染负荷比按大小排列，累计百分比大于 80%的污染物为主要污染物。将污染源等标污染负荷比按大小排列，累计百分比大于 80%的污染源为主要污染源。

$$K_i=\frac{P_m}{P} \quad (5)$$

$$K_j=\frac{P_n}{P} \quad (6)$$

式中：$K_i$——污染区内某种污染物的等标污染负荷比；

$K_j$——污染区内某污染源的等标污染负荷比；

$P_n$——某污染源内各污染物的等标污染负荷之和；

$P_m$——污染区各污染源内某种污染物的等标污染负荷之和；

$P$——污染区内所有污染源的等标污染负荷之和。

b）其他方法。对于存在环境健康风险且不能通过等标污染负荷法筛选的污染源和污染物，可根据研究目的，依据国家发布的有毒有害污染物名录或者环境与健康调查研究结果，结合研究地区实际情况，确定高环境健康风险污染源和特征污染物，将其纳入统计分析。

### 8.3.2 污染源监测数据分析

8.3.2.1 分析内容

污染源排放的特征污染物的检出情况、浓度水平及超标情况。

8.3.2.2 分析方法

a）检出情况。计算特征污染物的检出率（公式 7）或计算检出样品数与样品总数的比，来描述其检出情况。

检出率=测定结果高于方法检出限的样品数/样品总数×100% （7）

b）浓度水平。检验特征污染物浓度数据的分布，根据数据分布特征，依据 7.1 选取统计指标描述其平均水平。

c）变异程度。根据特征污染物浓度数据的分布特征，依据 7.1 选取统计指标描述其变异程度。

d）超标情况。选取污染物排放标准，计算特征污染物排放浓度的超标率（公式 8）或计算超标样品数与样品总数的比，描述其超标情况。

$$超标率=超过相应标准的样品数/样品总数\times 100\% \quad (8)$$

#### 8.3.3 污染源分布状况

根据污染源的空间位置信息，采用空间地图等展示污染区内污染源的分布状况。

## 9 环境暴露调查数据分析

### 9.1 统计分析目的

a）统计各环境介质中特征污染物浓度水平，判断调查区（污染区、对照区）的污染程度及污染物的时空分布特征；

b）进行不同调查区之间污染水平的差异比较，进一步明确污染影响范围及程度；

c）结合人群暴露参数估算人群暴露量。

### 9.2 数据特征

环境监测数据多呈非正态分布，具有严格时间或空间序列特征，时空相邻的数据具有更大可能的相似度，人为或自然原因可能引起瞬间或局部环境监测数据的变化。

### 9.3 统计分析内容及方法

#### 9.3.1 环境暴露水平描述

对环境空气、环境水体、土壤、室内空气、饮用水、农畜水产品、室内积尘等样本中特征污染物的检出情况、浓度水平、变异程度和超标情况进行统计描述。统计内容及指标包括：

a）检出情况。统计样本例数，并计算特征污染物的检出率；

b）浓度水平。检验数据分布类型，依据 7.1 选取统计指标描述其平均水平；

c）变异程度。根据特征污染物浓度数据分布特征，依据 7.1 选取统计指标描述其变异程度；

d）超标情况。有相关标准或参考值时，计算样本超标率和平均超标倍数；无相关标准或参考值时，计算实测平均浓度与当地背景值或对照区污染物平均浓度的比值；无背景值时可参考相关文献。

#### 9.3.2 环境暴露水平比较

依据 7.2 选取统计方法对不同调查区环境样本中特征污染物平均水平进行差异比较。

#### 9.3.3 暴露量估算

参见《环境污染物人群暴露评估技术指南》（HJ875-2017）。

## 10　人群健康调查数据分析

### 10.1　统计分析目的

a）比较特征污染物对人群的健康影响（包括特征污染物人体内负荷水平变化、生理功能或生化代谢变化、机体功能失调、发病及死亡等）在不同环境暴露水平之间的差异；

b）比较特征污染物的健康影响在不同人群之间的差异。

### 10.2　数据特征

人群健康状况往往与企业污染排放、调查区的环境质量状况存在一定的时空相关性，健康调查数据呈现组群间差异较大、群内差异较小的特点。

### 10.3　统计分析内容及方法

#### 10.3.1　人群基本情况

##### 10.3.1.1　分析内容

研究对象的人口学特征（如年龄、性别、民族、婚姻状况、文化程度等）、行为危险因素情况（吸烟、饮酒、户外活动习惯等）、职业暴露史、既往患病情况、家族史、就医行为等信息。

##### 10.3.1.2　分析方法

a）依据 7.1 选取指标进行统计描述；

b）依据 7.2 选取统计方法进行差异比较。

#### 10.3.2　特征污染物人体内负荷水平

##### 10.3.2.1　分析内容

调查区人群的特征污染物人体内负荷水平。

##### 10.3.2.2　分析方法

a）依据 7.1 选取统计指标对调查区人群的特征污染物人体内负荷水平的平均水平及变异程度进行统计描述；

b）统计样本例数、特征污染物人体内负荷水平正常例数和异常例数，依据 7.1 采用率（正常率、异常率、检出率等）或分数对特征污染物人体内负荷水平的正常情况、异常情况及检出情况进行统计描述；

c）依据 7.2 选取统计方法对调查区人群的特征污染物人体内负荷水平及正常率、异常率、检出率等进行差异比较。

#### 10.3.3　人群健康状况

##### 10.3.3.1　分析内容

调查区人群健康状况（如症状、体征、疾病、死亡等）的分布及其影响因素。

a）当对调查区人群的疾病（死亡）率进行描述时，计算总率；

b）当按人口学特征（人群的年龄、性别、职业、民族等）以及疾病种类描述疾病率时，计算专率。

c）当比较两组或几组人群的健康状况时，不应计算粗率，应按年龄、性别等可能影响健康状况的因素进行标准化，计算调整率。标准化计算的关键是选择统一的标准构成，选取标准构成的方法有三种：

1）选取有代表性的、较稳定的、数量较大的人群构成作为标准构成，如全国范围或全省范围的人口数据作为标准人口构成；

2）选择用于比较的各组例数合计作为标准人口构成；

3）从比较的各组中任选其一作为标准人口构成。

10.3.3.2　分析方法

a）症状和体征。统计各种症状体征的阳性例数，依据 7.1 选取率（阳性率）或分数对各种症状体征的异常情况进行统计描述，依据 7.2 选取统计方法对不同人群间的差异进行比较。

b）生理生化指标。依据 7.1 选取相应指标对不同人群生理生化指标的平均水平及变异程度进行统计描述；依据 7.1 选取率（正常率、异常率）或分数对不同人群生理生化水平进行统计描述，依据 7.2 选取统计方法对不同人群间的差异进行比较。

c）患病和死亡。统计不同疾病的新发病例数、患病例数、死亡例数，依据 7.1 选取相应的率（发病率、罹患率、患病率、死亡率）或分数对疾病的患病死亡情况进行统计描述，依据 7.2 选取统计方法对不同人群间的差异进行比较。

## 11　关联性分析

### 11.1　统计分析目的

通过对特定时点/时期、特定范围内开展的污染源调查、环境暴露调查、人群健康调查所获取的数据进行统计分析，探索某个或某几个环境因素对人群健康影响的可能性。

### 11.2　分析内容

a)比较污染区与对照区之间，或根据距离污染源远近划分的不同调查区之间环境暴露水平的差异，分析污染源对环境介质的影响范围及程度；

b）比较污染区与对照区之间，或不同环境暴露组别间人群健康水平的差异，分析环境因素与健康水平的关联性。

### 11.3　分析方法

#### 11.3.1　污染源与环境暴露

a）参照 9.3.2 的统计分析结果，比较污染区与对照区之间，或根据距离污染源远近划分的不同调查区之间的环境暴露水平的差异。

b）采用散点图描述特征污染物的浓度水平随与污染源距离的变化情况。若散点图呈现线性相关，则采用相关系数和线性回归模型定量描述距离与污染物浓度水平的相关程度，相关和回归模型选取原则参照 7.2。

c）采用统计地图标示污染源的地理位置，并采用空间插值等方法描述特征污染物浓度水平的分布情况，定性判断污染源与环境介质中特征污染物浓度水平在空间分布上的相关性。污染物浓度水平的空间插值常用方法包括克里金法、样条函数法、反距离插值法和趋势面法。通过比较各方法插值结果的交叉验证指标（包括平均误差、平均绝对误差、平均相对误差、均方根误差）选择误差相对较小的插值方法。

#### 11.3.2　环境暴露与人群健康

根据特征污染物和健康效应变量类型选取环境暴露与健康水平关联性分析方法。

a）健康数据为计量资料时，采用散点图描述环境暴露水平与人群健康水平的相关性，若散点图呈现线性相关关系，则依据 7.2 选取线性相关系数和线性回归模型定量描述相关程度。

b）健康数据为计数资料时，根据污染区与对照区的划分或环境暴露的不同水平分组，描述每个组对应的人群健康指标（如阳性率/异常率/患病率等），依据 7.2 选取 $\chi^2$ 检验、连续性校正 $\chi^2$ 检验或 Fisher 确切概率检验方法进行不同暴露水平组间人群健康水平的差异比较。采用 Logistic 回归模型定量分析环境暴露水平（自变量）与人群健康水平（因变量）的相关性。

c）某一环境暴露因素导致的健康效应可能长达数年，可根据实际情况选取既往的环境暴露水平与当前的健康效应水平进行关联性分析。

### 11.3.3 关联性判断

#### 11.3.3.1 散点图

散点图中 $Y$ 值随 $X$ 值增加而上升，则 $Y$ 与 $X$ 有正相关关系；$Y$ 值随 $X$ 值增加而下降，则 $Y$ 与 $X$ 有负相关关系；$Y$ 值与 $X$ 值增减无一定规律，或 $Y$ 值的变化不受 $X$ 值变化的影响，则 $Y$ 与 $X$ 不具有相关关系；$Y$ 值与 $X$ 值增减服从非直线规律，则 $Y$ 与 $X$ 无线性相关关系。

#### 11.3.3.2 相关系数

$0<r$（或 $r_s$）$<1$ 且 $P<\alpha$（一般取 0.05），则两变量呈显著正相关关系；$-1<r$（或 $r_s$）$<0$ 且 $P<\alpha$，则两变量呈显著负相关关系；$r$（或 $r_s$）$=0$ 且 $P<\alpha$，则统计学意义上两变量不相关。

#### 11.3.3.3 回归分析

对于线性回归和 Poisson 回归，回归模型的 $P<\alpha$，单个自变量的回归系数 $\beta_i$ 的 $P<\alpha$，则认为此自变量与因变量显著相关；若回归系数的 $P>\alpha$，认为此自变量与因变量无显著相关性。

对于 Logistic 回归，若环境暴露水平为连续型变量，判断方法同上；若环境暴露按不同水平分组，研究区的环境暴露水平高于对照区，且研究区相对于对照区的人群患病比值比（*OR* 值）$>1$ 且 $P<\alpha$（或 *OR* 值置信区间下限大于 1），则提示研究区内的环境污染与人群健康之间可能存在相关关系。

#### 11.3.3.4 样本统计量比较

如果研究区内特征污染物的浓度水平或人群特征污染物暴露量显著高于对照区（$P<\alpha$），相应的人体内负荷水平、生理生化指标异常率、症状体征损害指标和疾病患病情况也显著高于对照区（$P<\alpha$），则提示研究区内的环境污染与人群健康之间可能存在相关关系。

## 12 统计结果表达

通过统计表和统计图展示统计分析结果，并给予必要的说明，统计图和统计表的制作参见附录 A 和附录 B。数值修约依据 GB/T 8170 进行。

# 附　录　A
（资料性附录）
统计表编制原则和结构

## 1　编制原则

a）一张表只表达一个中心内容和一个主题。若内容过多，可分别制成若干张表。

b）主谓分明，层次清楚。统计表表达的是若干完整的文字语句，主谓语的位置要准确。定语部分放在标题内，主语放在表的左边作为横标目，谓语放在右边作为纵标目，横标目与纵标目交叉的格子放置数据，从左向右读，每一行便形成一个完整的句子。如表 A-1 中第一行可读为“叶菜在地区 A 的样本例数为 18，铅含量水平的中位数为 1.53mg/kg，超标率为 100%；在地区 B 的样本例数为 45，铅含量水平的中位数为 1.22mg/kg，超标率为 100%”。

c）数据表达规范，文字和线条尽量从简。

## 2　统计表结构

统计表可由标题、标目（包括横标目、纵标目）、线条、数字和备注 5 部分构成，示例见表 A-1。

a）标题。简明扼要地说明表的主要内容，包括时间、地点和研究内容，放在表的上方正中位置。如果有多张表格，标题前应加上标号。如果表中所有数据指标的度量衡单位一致，可以将其放于括号内置于标题后面。

b）标目。包括横标目和纵标目，有单位时需标明。横标目位于表的左侧，说明每一行数据的意义，纵标目位于表头右侧，说明每一列数据的意义。纵标目的总标目主要是对纵标目内容的概括，需要时设置。

c）线条。通常采用“三线表”格式，即顶线、底线、纵标目下的横线。若某些标目或数据需要分层，可用短横线分隔。

d）数字。用阿拉伯数字表示，同一指标小数点位数一致、位次对齐。表内不留空格，无数字用“-”表示，以备注的形式说明。若数字是“0”，则填写“0”。

e）备注。表中数据区需要插入文字或其他说明，可用阿拉伯数字或英文字母或符号（如“*”）等以上角标形式标出，将说明文字写在表格的下面。

**表 A-1　某调查地区夏季食物中铅含量水平** [1]

| 种类 | 地区 A | | | 地区 B | | |
|---|---|---|---|---|---|---|
| | 例数 | 中位数（mg/kg） | 超标率（%） | 例数 | 中位数（mg/kg） | 超标率（%） |
| 叶菜 | 18 | 1.53 | 100 | 45 | 1.22 | 100 |
| 土豆 | 10 | 0.85 | 100 | 20 | 0.79 | 100 |
| 鸡蛋 [2] | 0 | - | - | 28 | 0.13 | 23 |

[1] 《食品中污染物限量标准》（GB 2762-2017）：芸薹类和叶菜类蔬菜铅限量值为 0.3mg/kg；豆类蔬菜、薯类铅限量值为 0.2mg/kg，蛋及蛋制品（皮蛋、皮蛋肠除外）铅限量值为 0.2mg/kg。

[2] 未调查地区 A 鸡蛋中铅的含量。

# 附 录 B
（资料性附录）
统计图制作原则和结构

## 1 制作原则

a）根据资料的性质、分析目的选用适当的统计图。

b）一张图只表达一个中心内容和一个主题，即一个统计指标。

c）编制图形应注意准确、美观，图线粗细适当，定点准确，不同事物用不同线条（实线、虚线、点线）或颜色表示，给人以清晰的印象。

## 2 统计图结构

统计图通常由标题、图域、标目、图例和刻度 5 个部分组成，示例见图 B-1。

a）标题。简明扼要地说明资料的内容、时间和地点，置于图的下方正中位置并编号。

b）图域。即制图空间，除圆图外，一般用直角坐标系第一象限的位置表示图域，或者用长方形的框架表示。

c）标目。分为纵标目和横标目，表示纵轴和横轴数字刻度的意义，有度量衡单位时需标明。

d）图例。对图中不同颜色或图案代表的指标注释。图例通常放在图的右上角或图的正下方。

e）刻度。即纵轴与横轴上的坐标。刻度可在内侧或外侧，刻度数值按从小到大的顺序，纵轴由下向上，横轴由左向右。纵坐标原点必须从零开始。

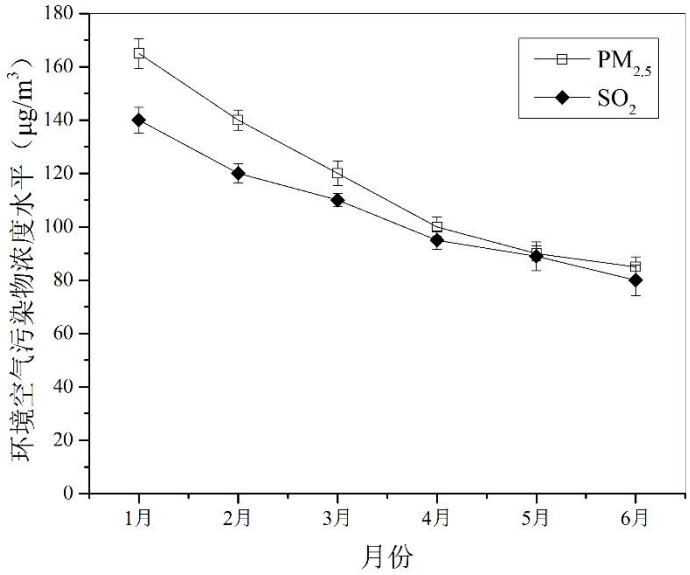


图 B-1 某地区环境空气污染物的月均浓度水平