

医学

统计方法

上海第一医学院卫生统计学教研室

医学统计方法

上海第一医学院卫生统计学教研组

上海科学技术出版社

内 容 提 要

本书主要是根据历年在教学、科研中应用数理统计方法的例子总结编写而成的一本实用的参考书。

书中除介绍目前医学中常用的统计方法以外，同时也介绍了一些多变量分析的内容，都作了深入浅出的叙述，为了便于非数学专业人员参考，对数理统计原理未作详细讨论。

本书可供医学、卫生、生物等科学研究人员、技术人员、医学院校教师及有关医务人员参考。

3486/16

医 学 统 计 方 法

上海第一医学院卫生统计学教研组

上海科学技术出版社出版

(上海瑞金二路450号)

新华书店上海发行所发行 上海中华印刷厂印刷

开本 787×1092 1/16 印张 20.75 字数 490,000

1979年8月第1版 1979年8月第1次印刷

印数 1—30,000

书号: 14119·1360 定价: 1.95元

前 言

为在本世纪内把我国建设成为现代化的社会主义伟大强国，广大医药卫生人员正在向科学技术现代化进军。为了更有效地进行医学科学实验，推动医药卫生工作不断发展，掌握一定的医学统计方法是很重要的。

为此目的，我们主要根据历年在教学、科研中所应用的统计方法，总结编写了这本重在实用的参考书。为了使内容浅显易懂，便于医务人员学习，尽可能多用直观性逻辑说明，以代替数理统计理论上的严格推导（一部分公式的逻辑性证明附在书末），多举一些实际的例子，以使读者掌握一些医学科学研究中常用的统计方法，供实验设计和数据分析时借鉴。

近年来医学统计方法有着很大的发展，对实验设计和分析方法都提出了新的要求，为了适应医药卫生人员医学科学研究工作的需要，在本书中我们除对一般统计方法、实验设计、非参数统计、Ridit 分析、序贯分析等作分章叙述外，还编写了多变量分析中的逐步回归分析与判别分析两章。

参加本书编写和核校人员有：

许世瑾 金丕焕 顾遂珏 陈启明 顾杏元 蔡文玮 詹绍康 龚幼龙 曹素华
严思刚等同志。

序贯分析一章主要参考本院药理教研组徐端正同志文稿编写而成。

由于我们的知识和经验有限，本书中缺点错误在所难免，恳切希望读者批评指正。

上海第一医学院卫生统计学教研组

一九七八年六月

目 录

前 言	
第 一 章 绪论	1
第一节 统计方法在医学科学研究及疾病防治工作中的作用	1
第二节 统计研究的对象	1
第三节 几个基本概念	2
第四节 正确的观点与思想方法	3
第 二 章 统计资料的来源和整理	5
第一节 统计资料的来源	5
第二节 统计资料的整理	7
第 三 章 平均数与标准差	11
第一节 算术平均数和几何平均数	11
第二节 中位数和百分位数	15
第三节 标准差	18
第 四 章 均数的抽样误差、可信限与 t 分布	24
第一节 均数的抽样误差	24
第二节 均数的可信限与 t 分布	25
第 五 章 t 检验	29
第一节 t 检验一般步骤	29
第二节 同一批对象实验(或处理)前后差别均数的显著性检验	30
第三节 两个样本均数差别的显著性检验	32
第四节 方差不齐时两样本均数差别的显著性检验(t' 检验)	34
第五节 两种检验与两类错误	35
第 六 章 方差分析(一)	37
第一节 方差分析的基本概念	37
第二节 完全随机设计资料的方差分析	37
第三节 随机区组设计资料的方差分析	41
第四节 各组均数间的相互比较	42
一、各组含量相等时的比较	42
二、各组含量不等时的比较	43
第五节 几个处理组均数与一个对照组均数比较	45
第六节 拉丁方设计资料的方差分析	46
第七节 方差齐性检验	48
一、各样本含量相等时检验方差齐性的步骤	49
二、各样本含量不等时检验方差齐性的步骤	50
第八节 近似 F 检验法	51
第九节 方差分析中的数据转换	53
第 七 章 方差分析(二)	57
第一节 2×2 析因实验设计数据的方差分析	57

第二节	“2 ³ 设计”或“2×2×2设计”资料分析	59
第三节	各组次级组含量不等的方差分析	62
一、	两因素各次级组的含量不等,但呈一定的比例	63
二、	次级组含量不等且不呈比例	64
第四节	不同模型资料的 F 检验	69
第八章	直线回归与相关	72
第一节	直线回归	72
第二节	相关	82
第九章	多元线性回归与曲线回归	87
第一节	多元线性回归	87
一、	二元线性回归分析	87
二、	复相关系数和偏相关系数	91
第二节	曲线回归	91
一、	概述	91
二、	医学上常用曲线的配合方法	93
三、	曲线配合的拟合度——相关指数	98
第十章	协方差分析	100
第一节	协方差分析法的意义和应用	100
第二节	完全随机化设计资料进行协方差分析的步骤	101
第三节	简化计算协方差分析方法	104
第十一章	相对数的分析	107
第一节	相对数	107
第二节	率的差别显著性检验	110
第三节	率的标准法	113
第十二章	卡方检验(一)	121
第一节	四格表资料的 χ^2 检验	121
第二节	行×列表的 χ^2 检验	123
第三节	四格表精确检验法	125
第四节	配对计数资料的 χ^2 检验	128
第十三章	卡方检验(二)	131
第一节	2× C 表 χ^2 的计算公式和 χ^2 值的分割	131
第二节	组内分组资料的卡方检验	133
第三节	内部构成不同的两个率的差别显著性检验	135
一、	加权 χ^2 检验法(Cochran 检验)	135
二、	χ^2 值相加法	138
三、	χ 值相加法	138
第四节	分表理论数合并法	139
一、	2×2表理论数合并法	139
二、	行×列表理论数合并法	140
第五节	行×列表的分割	141
一、	2× b 表的分割	142
二、	行×列表的分割	143

第六节 $2 \times C$ 表线性回归的显著性检验	144
一、按性质分组的资料的百分率(比)趋势的显著性检验	145
二、按数量分组的资料的百分率趋势的显著性检验	146
三、当百分率很小(或很大)时的趋势的显著性检验	147
四、药物疗效比较	147
第十四章 二项分布与普哇松分布	149
第一节 二项分布	149
第二节 普哇松分布	153
第十五章 非参数统计	160
第一节 成对资料的显著性检验	160
一、符号检验法	160
二、符号等级检验法(Wilcoxon 法)	162
第二节 成组比较资料显著性检验	163
一、两样本等级和检验法(Wilcoxon, Mann and Whitney 法)	163
二、中位数检验法	164
第三节 完全随机化设计资料检验	165
一、 H 检验法(Kruskal and Wallis 法)	165
二、中位数检验法	166
第四节 随机区组资料的非参数统计法	167
一、 M 检验法(Friedman 法)	167
二、中位数检验法	168
第五节 等级相关(Spearman 法)	169
第十六章 Ridit 分析	171
第一节 样本与总体比较	171
第二节 两组平均 Ridit 比较	175
第三节 多组平均 Ridit 差别的显著性检验	176
第十七章 序贯检验	179
第一节 序贯检验的意义	179
第二节 单一处理效果的检验——单向检验	180
第三节 两种处理的配对比较——双向检验	184
第四节 闭锁型序贯检验	188
第十八章 逐步回归分析	192
第一节 基本概念	192
第二节 计算方法和步骤	192
第三节 逐步回归分析注意事项	197
第十九章 判别分析	199
第一节 基本概念	199
第二节 两类判别,两个指标	200
第三节 三个指标的两类判别	204
第二十章 半数效量	206
第一节 基本概念	206
第二节 目测法	207

第三节	加权直线回归法	209
第四节	Kärber 氏法(寇氏法)	214
第五节	序贯法	215
第六节	半数效量的有关应用	216
第二十一章	生存率统计	218
第一节	直接法	218
第二节	寿命表法	220
第二十二章	研究设计	224
第一节	研究设计的基本内容	224
第二节	不同场合的研究设计	227
一、	调查研究设计	227
二、	临床研究设计	227
三、	动物实验设计	228
第三节	研究设计的类型	228
一、	完全随机化设计	228
二、	随机区组设计	229
三、	拉丁方设计	231
四、	序贯设计	234
五、	其他设计	234
第二十三章	抽样设计	235
第一节	单纯随机抽样	235
第二节	机械抽样(系统抽样)	236
第三节	分层抽样	236
第四节	整群抽样	238
第二十四章	样本大小	241
第一节	对平均数作抽样调查时的样本大小	241
第二节	对率作抽样调查时的样本大小	242
第三节	对样本均数与总体均数的差别作显著性检验时所需样本大小	243
第四节	成组比较即两个样本均数的差别作显著性检验时所需样本大小	244
第五节	成对比较(或自身前后对比)作显著性检验时所需样本大小	245
第六节	对两个率的差别显著性检验所需样本大小	246
第二十五章	统计表与统计图	247
第一节	统计表	247
第二节	统计图	249
附录一	统计用表	256
附表 1	正态分布表	256
2	t 值表	257
3	F 值表(双侧检验,方差齐性检验用)	258
4-1	F 值表(单侧检验,方差分析用) $P=0.05$	260
4-2	F 值表(单侧检验,方差分析用) $P=0.01$	261
5	各均数间相差显著时所需之 Q 值	262
6-1	Dunnnett t 表(单侧检验)	263

6-2	Dunnett t 表(双侧检验)	264
7	相当于百分比之角度($A = \sin^{-1} \sqrt{p}$)	265
8	r 值表	267
9	由 r 转 Z 值表	268
10	χ^2 值表	268
11	$\lg n!$ 数值表	269
12	四格表的显著性检验用表(C 值表)	270
13-1	二项分布中 95% 可信限(百分比)	278
13-2	二项分布中 99% 可信限(百分比)	279
14	普哇松分布用的上下可信限	280
15	符号等级检验表(Wilcoxon 成对比较用)	280
16	等级总和数临界值(双侧检验)	281
17	H 值与概率对照表	282
18	计数资料单向序贯检验边界系数表	284
19	M 值的界限值($P = 0.05$)	286
20	等级相关系数的显著性界限表	286
21	计量资料单向序贯检验 $\frac{\theta_1 - \theta_2}{\sigma}$ 值表	287
22	配对计数资料序贯检验边界系数表	288
23	配对计量资料序贯检验边界系数表	288
24	计数资料闭锁型序贯检验边界(U, L)和中间线(M, M')坐标	288
25	计量资料序贯检验闭锁线坐标 n', y'_n 数值	289
26	百分率与概率单位对照表	290
27	概率单位与权重系数对照表	290
28	作业用概率单位之极小值、极大值及全距	291
29	随机数目表	292
30-1	对平均数作抽样调查, S/δ 取不同的数值时所需样本大小 n ($\alpha = 0.05$)	293
30-2	对平均数作抽样调查, S/δ 取不同的数值时所需样本大小 n ($\alpha = 0.01$)	293
31-1	对率作抽样调查时所需样本大小 n ($\alpha = 0.05$)	294
31-2	对率作抽样调查时所需样本大小 n ($\alpha = 0.01$)	294
32	对样本均数与总体均数的差别作显著性检验时所需样本大小 n	295
33	对两个样本均数的差别作显著性检验时所需样本大小 n	296
34-1	对两个率的差别作显著性检验时每组所需样本大小 n ($\alpha = 0.05, \beta = 0.10$)	297
34-2	对两个率的差别作显著性检验时每组所需样本大小 n ($\alpha = 0.01, \beta = 0.05$)	297
35	平方表	298
36	平方根表(\sqrt{n} 和 $\sqrt{10n}$)	300
37	常用对数表	304
38	反对数表	306
附录二	统计符号表	308
附录三	参考文献	311
附录四	常用统计公式证明	312
附录五	希腊字母表	318
附录六	汉英名词对照	319

第一章 绪 论

第一节 统计方法在医学科学研究及疾病防治工作中的作用

统计方法是认识社会和自然现象数量特征的重要工具。正确的统计分析能够帮助我们正确认识事物客观存在的规律性。我们做任何工作,都必须掌握基本情况,做到胸中有数,才能有的放矢地开展,从而提高工作质量。进行医学科研和疾病防治工作也不例外。无论在基础医学、临床医学和预防医学各个方面的科学研究以及防治工作计划的拟订和成绩的正确评价,都必须有计划地收集资料并进行合理的统计分析。例如通过病史分析,进行某种中草药疗效的研究和各种疗法疗效的比较;通过流行病学调查,研究外界环境因素、生活因素、劳动因素等与疾病的关系;通过疾病登记资料的分析,了解哪些疾病是常见病、多发病以及它们的发生发展规律,哪些是危害健康最严重的疾病;通过历年资料的对比,可以了解在防治工作中已取得了哪些成绩,还有哪些差距,今后努力的方向是什么,如何制订有效的防治措施的计划,等等。因此,我们医务工作者必须学习和掌握统计方法。在日常工作中要认真做好一切记录,有计划地进行调查和检验。积累了一定的资料以后,就可以运用统计方法将资料科学地进行整理分析。透过众多的偶然的次要的因素阐明事物客观存在的规律性,辨别事物间在数量上的差别是否仅是偶然现象,从而作出比较正确的结论应用于实际,并且针对工作中发现的问题,制订措施,以提高防治工作的质量,更好地为广大劳动人民服务。

第二节 统计研究的对象

统计所要研究的对象是有变异的东西。在相同的对象中往往存在着变异。例如同为健康人,即使是同性别、同年龄,他们的身长、体重、血压、脉搏、体温、红细胞、白细胞等数值都会有所不同。同为一种病人,病情轻重可以有所不同。对病情相同的病人,用同一疗法治疗,有的治愈,有的没有治愈。治愈的病例其病程也有长有短。以上这一些都是变异的例子。如果研究对象内各个个体都完全相同,没有什么变异,如某一药物的化学结构等,我们只要分析一个个体,就可以了解总体。象这种没有变异的东西,不是统计研究的对象。

由于事物之有变异,统计必须在一定数量观察的基础上进行。统计不是孤立地研究各种现象,而是通过一定数量的观察,从这些现象里研究事物间的相互关系,阐明事物客观存在的规律。由于统计研究对象存在着变异,变异的出现是由于许多内外因素偶然性的配合所致,因此,统计研究的各种现象的表现结果是一种随机事件。随机事件是指一次试验不确定而在一定数量重复条件下呈现出统计规律性来的事件。科学研究的目的在于阐明客观存在的规律性,从而通过它们对同类事物加以估计和预测,以便应用于实际。所以统计须在一定数量观察的基础上进行研究。

统计研究还必须注意所研究的事物在性质上是否相同(统计上称为“同质”)。例如,研

究某种新药对肝癌的疗效,所有研究对象都必须是确诊为肝癌的病人,不可包括疑似病人或根本不是肝癌病人在内。务使研究对象是同质的,否则研究结果就不能反映这个新药对肝癌的疗效的客观规律。判断事物是否同质,需要结合业务知识和研究目的来考虑。例如我们要了解某地区人群的肺结核患病严重程度,利用医院X线检查资料计算肺结核阳性百分率。医院X线检查资料看起来是同质的,但实际上医院资料有两种来源:一种是门诊和住院肺结核疑似病人的检查资料,一种是人群胸部X线检验的普查资料。这两种资料显然不是同质的。如果我们要了解某一人群的肺结核患病率,我们所需要的是这一人群的普查资料而不是疑似病人的检验资料。

第三节 几个基本概念

一、总体与样本

总体是同质的个体所构成的集体。总体所包含的个体数可以为无穷大,而且往往只是设想的或是抽象的。例如就宫颈癌病人来说,所有宫颈癌病人都具有患宫颈癌这个同质的特征,是一个总体。每一个宫颈癌病人叫做个体。但我们要研究宫颈癌的规律,事实上并不能将宫颈癌病人总体都观察到,而只能对一部分个体来进行观察。这种从总体中取出部分个体的过程称为“抽样”。所抽得的部分就称为样本。在一个样本里可以含有不同的个体数。这个样本所包含的个体数目称为样本含量。

在科研工作中,手头上所掌握的资料,经常只是总体的一小部分,只是一个样本。但科学研究的目的,不能满足于依据一个样本所得到的结果,而期望知道总体的情况。例如用某种新药治疗高血压病人100例,有效的60例,有效率为60%。但我们所观察到的只是含量为100的一个样本。如果用这个新药治疗其他高血压病人,是否也会获得同样的效果呢?这就需要以样本的疗效来估测总体的疗效。如何正确地从样本来推测总体,这是统计所要解决的问题。

二、抽样误差

由于总体中的各个个体存在变异,因此在同一总体中随机抽取若干个含量(个体数)相同的样本,各样本均数(或率)相互间会有所不同,这些样本间的差异同时反映了样本与总体间的差异。这种由于抽样而引起的差异,在统计上称为抽样误差。

一般来说,样本越大,则抽样误差越小,越和总体的情况相接近,正确性越高,换言之,越能阐明事物的规律性。但在实际工作中,我们往往不可能观察太多的对象,而只能从一个较小的样本资料来进行分析研究,计算出一系列的统计指标。样本越小,抽样误差相应地越大,如果仅凭数字的表面值进行判断,不考虑偶然性问题,往往会导致错误的结论。但是,如果能够正确地应用“概率论”*所导出的一些统计处理方法,我们就能辨别在实验研究中哪一些是有实际意义的,哪一些可能仅是由于偶然性所造成的,从而有助于得出一个正确的结论。这在医学实践上甚为重要。关于这一类的统计处理(或检验)方法甚多,以后将择要介绍。

三、随机化

随机化是抽样研究和抽样分配时十分重要的原则。从研究对象的总体中抽取样本进行

* “概率论”是一门研究随机事件的数量规律的学科。

研究时,为了使样本对总体有较好的代表性,并使其抽样误差的大小可以用统计方法来估计(因为一般数理统计所提供的方法是在随机抽样的基础上推演而来的),必须遵守随机化原则,就是说在抽取样本以前要使总体中每个单位有同等的被抽取的机会。同样的,在动物实验中,把动物分配成为几组,也必须用随机方法将各个动物有同等的机会被分配到各组去,这样就不至于人为地造成各组间动物的不齐全。随机化的具体方法有抽签法和用随机数字表等(见第23章)。

四、概率,亦称机率、或然率

它是反映某一事件发生的可能性大小的量。例如我们说某药治疗某病的治愈率为80%(或0.8)。这个数值说明某药治愈某病的可能性,也就是说某药治愈该病的概率估计为0.8。统计上常用符号 P 来表示概率。必然发生的事件概率为1。不可能发生的事件概率为0。一般地概率范围是在0与1之间。概率越接近1,表明发生的可能性越大。概率越接近0,发生的可能性越小。我们在医学科研论著中常见到 $P \leq 0.05$,或 $P \leq 0.01$,前者表示事件发生的可能性等于或小于0.05,后者表示事件发生的可能性等于或小于0.01,作为事物差别有显著意义与有非常显著意义的界限(详见第5章)。

五、统计量与参数

在科研工作中,为了掌握研究对象的某些性质,我们对研究对象作了若干次观察,依据观察值所定出的量,称为“统计量”。例如为了了解健康成年男子每立方毫米血液中的白细胞数而进行检验,由所检得的一系列数值算出一个平均数,这个平均数(即样本平均数)是一个“统计量”。样本标准差和样本的率也是统计量。从这些统计量可以估计总体平均数、总体标准差和总体的率。这些总体的统计指标数值称为参数。

第四节 正确的观点与思想方法

应用统计方法于医学科学研究时,必须以唯物辩证法为指导,以医学科学为基础,有正确的观点与思想方法才能发挥它的作用。“统计无用”和“统计万能”的论点都是错误的。

有些人认为只凭常识判断,仅凭数字的表面值就可以看出实验研究各组间的差别,不需要什么统计处理。这种没有考虑到抽样误差问题的判断往往是错误的。当实验资料数量多而两组间的差别相当大的时候,凭常识判断和统计处理的结论可能是一致的;但如果实验资料少,特别是在两组差别不太大的情况下,如不考虑机遇问题,仅凭常识判断下结论,那就难免要发生错误。我们看到许多与医学有关的各种实验研究报告,大多能够正确地运用统计方法、合理地分析资料、做出的结论具有很强的说服力;但也有不少仅凭数字的表面值就匆匆地下了断语,一经统计处理以后,所得的结论恰恰相反。例如某些药品并没有什么特殊疗效或预防效果,但被错误地认为有效,并且推广应用,以致造成一定的损失。这充分说明科学研究需要正确的统计方法处理。

正确地运用统计方法,能够帮助我们正确认识客观事物,阐明事物固有的规律,从而把感性认识提高到理性认识,用于指导医学实践。但是统计并不是万能的。它决不能改变事物的本来面目把原不存在的规律“创造”出来。有些人在进行科学研究以前没有充分考虑,收集了一些杂乱的不正确的资料,而希望用统计方法来予以弥补,对这种情况统计是无能为力的。有些人甚至只凭主观愿望出发,滥用或乱用统计方法去凑合预定的结论,这是违背唯

物论的认识论的,玩弄数字的行为,是非常错误的。医学统计和其他统计一样,只能阐明规律而不能“创造”规律。只有在唯物辩证法思想指导下,有计划地收集资料,正确、合理地整理资料和分析资料,才能得出符合事物客观规律的结论,才能用于指导医学实践,更好地为提高劳动人民健康水平、为社会主义革命和社会主义建设事业服务。

第二章 统计资料的来源和整理

统计工作一般分为三个步骤：收集资料、整理资料和分析资料。收集资料是进行统计工作的第一步，也是最重要的一步。如果收集资料的计划不周密，原始记录不正确，往往会造成整理分析的困难，甚至得出错误的结论，而这些缺点是难以用统计分析来补救的。因此，在收集资料以前必须根据研究目的深思熟虑，不仅要考虑到研究的内容和项目，而且还要考虑到如何选择研究对象以及用何种方法进行统计分析，以使用较少的人力物力得出较多的科学结论，这就是一项研究工作的设计问题。第二步是将收集来的资料进行科学加工、分组归纳汇总，使资料系统化，即整理资料。第三步是将整理汇总的资料，计算相应的统计指标和绘制必要的统计图表以对研究对象的特征、规律等作综合的说明，即资料分析。统计工作三个步骤是相互联系的，每一步骤产生错误，都会影响最后的分析及结论的正确性。

第一节 统计资料的来源

医学统计资料的来源主要有三方面：一、统计报表。二、医疗卫生工作原始记录和报告卡。三、专题调查或实验。

一、统计报表

医疗卫生工作报表如卫生工作基本情况年报表、医院工作年报表、传染病年(月)报表、疫情旬(月、季)报表等。根据国家规定的报告制度，由医疗卫生机构定期逐级上报。统计报表是了解居民健康状况的基本资料，为拟订医疗卫生工作计划与措施、检查、总结工作提供科学的依据，也是科学研究的主要资料。

二、日常医疗卫生工作记录和报告卡

医疗卫生工作的记录有门诊病历、住院病历、健康检查记录、医学检验记录等。


各种报告卡有恶性肿瘤发病或死亡报告卡、传染病报告卡、职业病报告卡、出生报告单、死亡报告单等。

为了使这些资料可以供医学科学研究使用，必须注意资料的完整性与正确性。在收集这方面资料时，经常会出现重复、遗漏或项目填写不清等情况，特别是报告卡很易重复。所以，对这些资料要经常检查与核对，发现错误及时纠正。医务卫生人员必须认真填写这类资料，从思想上认识原始资料记录的重要性。

医院病史是诊治疾病的工作记录，是实践经验的积累，也是研究疾病病情和疗效的基本资料，非常重要。但病史资料由于种种原因往往存在挑选性或其他缺点，在应用中要注意以下几个问题：

1. 医院病史可能具有挑选性 一般来说，大医院的住院病例往往多是病情较重的或疑难的病例，和小医院的有所不同，因此，用以代表这种疾病病情的一般情况是不全面的。用统计的术语来说，这是有挑选性的样本，它只能代表具有同样病情的总体。以不同医院的资料作对比时(如研究治疗效果等)应注意这个问题。

表 2-2 卡片举例

肺部X线缩影片检查记录卡	
	姓名 _____ 性别 _____ 实足年龄 _____
	地址 _____ 区 _____ 路 _____ 号 _____
	检查结果 _____
	检查日期 _____
	团体编号 _____ 检查号 _____
	检查者 _____

第二节 统计资料的整理

从各种统计报表、原始记录及专题调查所得到的大量原始资料,是分散的资料。要了解事物总的特征和发展情况,必须对这些原始资料进行科学的分组归纳,使资料系统化,便于进一步统计分析以反映被研究事物的规律性,这个过程称统计资料的整理。整理工作可分为检查资料、设计分组、拟整理表和归纳四个步骤。

一、检查资料

收集的资料在没有整理分析前称为原始资料。在整理时首先要对原始资料进行详细检查,主要从下面三方面进行:

1. 资料的完整性 检查原始记录是否有遗漏或重复,各项目是否均已填齐。
2. 资料的正确性 检查所有项目填写是否正确,各项目之间有无矛盾,数字有无不合理之处。如研究疗效时,治愈和好转必须严格区分,好转病例不能作为治愈,否则算得的治愈率也不正确。

对有错误、相互矛盾、或重复、遗漏的资料,应当作合理的修正、剔除和补充,必要时进行复查。在专题调查时,检查资料最好在调查时就进行,边调查边检查,以便及时改正错误。在实验工作中随时要核对记录的正确性、完整性。

二、设计分组

原始资料经过检查后即可进行分组整理。所谓分组,是根据各项目的特性,将性质相同(同质)的资料归纳到一起,使资料系统化,才能反映事物的本质。因此,在整理分组时,必须坚持“同质”的原则,才能显示资料内部的规律性,得出正确的结论。

分组方法:

1. 质量分组 即按事物的性质或类型来分组,如疾病分类、死因分类;人口按性别、职业、工种等分组;疗效按治愈、好转、死亡分组;粪检虫卵按阳性、阴性分组等。不同质的资料不能并在一组。如研究心脏病与年龄有无关系时,不能将风湿性心脏病与高血压心脏病并在一组,因为这两组心脏病病人年龄分布不同,并在一组不能反映资料内部的规律性。
2. 数量分组 在质量分组的基础上再按数量(变量值)的大小来分组。如年龄、血压、脉搏、时间、红(白)细胞数等。

数量分组的多少,决定于资料的性质和资料的多少。分组过多和过少都不好,分组过多则规律性不能显著突出,分组过少又会隐蔽资料的特点。例如研究居民一般发病率时,年龄

分组不应以 0—19 岁为一组。因为这一组中包括了婴儿、幼儿和青少年，他们发病情况不同，如分在一组，就看不出各年龄组发病特点。又如为了研究流行性乙型脑炎各年龄的发病情况，因患者以儿童时期比较多，则要求把儿童时期的年龄分得细些，以 1 岁一组为宜，而 10 岁以上可分得粗些，可分成 5 岁或 10 岁一组。所以分组多少需根据资料性质而定。

为了便于进行资料间相互对比，还必须注意到习惯分组法。如年龄分组，习惯上分成 5 岁或 10 岁一组，分组时由 0 岁开始，分成 0—4 岁（指刚出生婴儿到未满 5 足岁者都包括在这一组内），5—9 岁（指刚满 5 足岁至未满 10 足岁），10—14 岁……等。在儿童年龄分组中如需分成 1 岁一组，可分为 0—, 1—, 2—, ……, 9—, (0—指刚出生婴儿至未满一足岁，余类推)。如分成 3 岁一组即分成 0—2 岁, 3—5 岁……等，则与其他资料作对比时就会产生困难。

此外，数量分组的界限要清楚。例如 0—5, 5—10, 10—15, ……，这种分组界限不清，5 究竟在 0—5 组内还是在 5—10 组内，容易使人误解，应写成 0—4, 5—9, 10—14, ……，或简写成 0—, 5—, 10—, ……，“0—”指从 0 起至未满 5 止。分组写法举例如下：

正确写法		错误写法
0—4	或 0—	0—5
5—9	5—	5—10
10—14	10—	10—15
15—19	15—	15—20
⋮	⋮	⋮
80 岁及以上	80 岁及以上	80 岁及以上

三、拟整理表

整理表是用于原始资料整理归组的一种表格。分组决定后便可根据研究目的和要求拟订整理表。在拟订时应注意各项目之间的相互关系，把关系密切的项目放在一个表内，使相互关系表达出来。例如根据上节所举的肺部 X 线缩影片检查记录拟整理表，如要从表里得到不同性别和年龄组肺部缩影片检查结果的人数，表 2-3 的设计是好的，我们可以从表中得到所需要的数字。而表 2-4 是设计不良的，因为性别与年龄孤立地分组，没有结合起来，不能从表中得到不同性别与年龄组检查结果的数字。

表 2-3 某地居民肺部 X 线缩影片检查结果整理表

年龄组(岁)	男		女		合 计	
	检查人数	阳性人数	检查人数	阳性人数	检查人数	阳性人数
0—						
5—						
10—						
15—						
⋮						
70 岁及以上						
不 详						
合 计						