基于SPSS的
健康数据分析与应用

ISBN 978-7-5690-8276-0

定价：58.00元

四川大学出版社
SICHUAN UNIVERSITY PRESS

基于SPSS的健康数据分析与应用

主编 许华 刘敏 景燕敏

四川大学出版社
SICHUAN UNIVERSITY PRESS

颐养有方 | 智慧康养系列新形态教材

基于SPSS的
健康数据分析与应用

主审 王英博

主编 许 华 刘 敏 景燕敏

四川大学出版社
SICHUAN UNIVERSITY PRESS

颐养有方 | 智慧康养系列新形态教材

基于SPSS的健康数据分析与应用

主　编　许　华　刘　敏　景燕敏

副主编　申文静　班曼丽　吴　双　张　猛

参　编　刘　娜　董雯倩　陶丙岭　郭晓艳　路梓艺

主　审　王英博

四川大学出版社
SICHUAN UNIVERSITY PRESS

图书在版编目（CIP）数据

基于SPSS的健康数据分析与应用 / 许华，刘敏，景燕敏主编．-- 成都 ：四川大学出版社，2026. 4.
ISBN 978-7-5690-8276-0

Ⅰ．R197.1-39

中国国家版本馆CIP数据核字第20254XN646号

书　　名：基于SPSS的健康数据分析与应用
　　Jiyu SPSS de Jiankang Shuju Fenxi yu Yingyong
主　　编：许　华　刘　敏　景燕敏

选题策划：庞国伟　王　睿
责任编辑：周维彬
责任校对：胡晓燕
装帧设计：黄燕美
责任印制：李金兰

出版发行：四川大学出版社有限责任公司
　　地址：成都市一环路南一段24号（610065）
　　电话：（028）85408311（发行部）、85400276（总编室）
　　电子邮箱：scupress@vip.163.com
　　网址：https://press.scu.edu.cn
印前制作：华腾教育排版中心
印刷装订：河北龙大印务有限公司

成品尺寸：202mm×278mm
印　　张：17.5
字　　数：468千字

版　　次：2026年4月 第1版
印　　次：2026年4月 第1次印刷
印　　数：1-3030册
定　　价：58.00元

本社图书如有印装质量问题，请联系发行部调换

四川大学出版社
微信公众号

前言

党的二十大报告指出，“促进优质医疗资源扩容和区域均衡布局，坚持预防为主，加强重大慢性病健康管理，提高基层防病治病和健康管理能力。”

在数字化浪潮席卷全球的当下，健康数据作为国家重要战略资源，正以前所未有的速度重塑医疗健康领域的发展格局。从个体日常健康监测到大规模公共卫生决策，从精准医疗临床实践到生物医药科研创新，健康数据统计分析已成为贯穿健康中国建设的核心纽带。本书立足时代需求，系统梳理健康数据统计分析的理论框架与实践路径，旨在为医疗卫生工作者、科研人员及相关专业学生，提供兼具科学性与实用性的学习指南。

《“健康中国 2030”规划纲要》明确提出“推进健康医疗大数据应用”的战略部署，将健康数据统计分析纳入国家健康治理体系的核心环节。《国务院办公厅关于促进“互联网＋医疗健康”发展的意见》进一步要求“健全‘互联网＋医疗健康’标准体系”，为数据的规范化采集、共享与分析提供制度保障。2023 年公布的《卫生健康统计工作管理办法》更是从法律层面明确了健康数据统计的责任主体、质量要求与应用规范，构建起国家统筹、地方协同、全员参与的统计工作格局。这些政策形成了从顶层设计到基层落实的完整链条，推动健康数据分析从技术探索走向制度化实践。

本书组建了由统计学、医学、公共卫生领域专家构成的跨学科团队，形成多层次的专业支撑体系。统计学专家在抽样方法设计、模型构建等方面提供理论指导，确保统计推断的严谨性；预防医学专家从流行病学视角解读数据关联，提升分析结果的公共卫生价值；三甲医院临床专家结合诊疗实践，明确数据指标的医学意义，避免统计分析与临床需求脱节。团队通过定期召开专家论证会、案例研讨会等形式，已解决数据标准化、数据隐私保护与价值挖掘的平衡等关键问题，形成多项具有实践价值的分析模型。

本书在内容编写上具有以下特色。

1. 模块化设计，结构清晰

本书采用科学的模块化设计，将健康数据统计分析的核心内容划分为三大模块——“健康数据采集与预处理”“健康数据分析”与“研究设计”。每个模块进一步细分为多个项目，并在项目中设置具体任务。这种层次分明的结构设计，使读者能够循序渐进地掌握知识体系，从基础概念到高级应用逐步深入，确保学习过程的系统性和逻辑性。同时，模块化的编排方式便于教师根据教学需求灵活调整授课顺序，也便于读者根据自身需求选择性学习重点内容。

2. 理论与实践结合，注重实操性

本书不仅系统讲解统计学的基本概念、原理和方法（如均值、标准差、假设检验、回归分析等），还紧密结合 SPSS 软件操作，提供详细的步骤指导和截图示例，确保读者能够将理论转化为实践。此外，每个任务均设有“实战演练”环节，要求读者基于真实数据集完成分析任务，如建立心理健康数据库、筛选特定人群数据等，从而巩固所学知识并提升实际应用能力。

3. 案例驱动，贴近实际应用

本书采用案例驱动的教学模式，通过真实场景的任务设计，将抽象的理论知识转化为具体的实践应用。每个学习任务都紧密结合公共卫生监测、临床诊疗评估、健康管理实践等专业领域的实际需求，让学习过程更具针对性和现实意义。案例化教学设计不仅注重知识传授，更强调培养读者的专业思维方式与问题解决能力。在学习过程中，读者需要像真正的健康数据分析师一样思考，从数据收集、清洗到分析、解释的每个环节都要考虑实际应用场景的特殊要求。这种沉浸式学习体验，有助于读者构建系统的知识框架，同时培养其将统计方法灵活运用于不同专业情境的能力，为未来职业发展奠定坚实基础。

4. 内容全面，涵盖健康数据分析全流程

本书系统构建了健康数据分析的完整知识体系，全面覆盖从数据源头到研究结论的全过程方法论。在内容架构上，首先深入探讨各类健康数据的获取途径与质量控制标准，继而详细阐释数据预处理的核心技术与规范流程。在核心分析部分，不仅包含基础统计描述方法，更系统介绍各类推断统计技术及其适用条件。最后延伸至研究方案设计的科学原理与方法学比较，完整呈现健康数据分析的学术逻辑与实践路径。这种循序渐进的内容编排，既保证了各知识模块的专业深度，又实现了不同环节间的有机衔接，能帮助读者建立系统化的分析思维，具备从原始数据到科学发现的完整研究能力。本书的全流程知识整合设计，有助于培养读者开展健康领域实证研究的综合素养。

健康数据分析正处于机遇与挑战并存的发展阶段。随着 5G、人工智能等技术的深度应用，健康数据的规模将持续扩大、维度将不断丰富，这既为精准健康管理提供了可能，也对统计分

析能力提出了更高要求。本书旨在搭建理论与实践之间的桥梁，帮助读者掌握科学的分析方法，在保障数据安全的前提下，充分释放健康数据的价值，为推进健康中国建设贡献专业力量。期待广大读者在使用过程中提出宝贵意见，共同推动健康数据统计分析学科的发展与完善。

本书由德州职业技术学院许华、刘敏、景燕敏任主编；由青岛工学院申文静，德州职业技术学院班曼丽、吴双、张猛任副主编；由德州职业技术学院刘娜、董雯倩、陶丙岭，桓台县中医院郭晓艳和山东大学齐鲁医院德州医院路梓艺任参编；由德州职业技术学院王英博任主审。

由于编者水平有限，书中难免存在不足之处，敬请广大读者批评指正。

编　者

目录

模块一

健康数据采集与预处理

项目二

健康数据清洗与筛选

项目导读

随着信息技术的发展，健康数据库在医疗保健领域发挥着越来越重要的作用。本项目将带领读者学习健康数据库中数据清洗与筛选的方法和技巧，通过实际案例操作，帮助读者掌握健康数据高效清洗与筛选的实操流程，为后续的数据分析和决策提供可靠依据。

教学目标

知识目标

理解健康数据库中常见的数据问题类型，如缺失值、异常值、错误值等。

掌握健康数据库数据清洗与筛选的概念、方法和工具。

能力目标

能够根据实际需求，运用合适的筛选条件对健康数据库中的数据进行筛选。

能够运用恰当的方法对健康数据库中的缺失值、异常值等进行预处理，提高数据质量。

能够运用数据清洗、转换等技术对健康数据库中的数据进行预处理。

素质目标

培养严谨的科学态度和数据意识，认识到数据质量对健康研究的重要性，确保在数据处理过程中操作认真、准确。

提升解决问题的能力，面对复杂的数据问题时，能够冷静分析，运用所学知识和技能寻找合适的解决方法。

增强团队协作精神，在项目实施过程中，能够与他人有效沟通、协作，共同完成数据清洗与筛选任务。

案例导入

本模块项目一建立了“老年人慢性病健康数据库”，但在实际的数据收集和整理过程中，数据可能存在不完整、不准确、不一致等质量问题，还要聚焦于特定的慢性病患者群体，排除不符合研究要求或存在错误的数据记录，从而提高分析的准确性和效率。因此，需要对数据进行系统性清洗与筛选。

任务一 清洗健康数据

任务描述


微课
数据清洗 1

在项目一任务二中，已经利用收集的数据建立了“老年人慢性病健康数据库”。本任务需要对“老年人慢性病健康数据库”中的数据进行清洗，包括处理缺

失值、异常值和错误值等，以提高数据质量，为后续的健康数据分析和建模奠定基础。

任务分析

对于“老年人慢性病健康数据库”中可能存在的问题，我们需要运用合适的方法对缺失值进行填补，对异常值进行修正或剔除，对错误值进行更正，以确保数据的质量。

知识链接

一、数据质量评估指标

（一）准确性

考查数据是否真实、可靠地反映了实际情况，需检查数据是否存在错误值、异常值或不符合逻辑的数据。例如，年龄不应为负数，血压值应在正常生理范围内。

（二）完整性

衡量数据集中应有数据的完整程度，即数据是否缺失。可以列出数据集中应包含的所有变量和记录，检查是否都已存在。

（三）一致性

判断数据在同一数据集中不同部分或不同数据源之间是否逻辑一致。

（四）唯一性

确保数据集中不存在重复记录，可通过唯一标识符（如身份证号）或多个关键变量（如姓名、性别、出生日期等）的组合来检测重复记录。

（五）时效性

确认数据反映的是否为最新或在规定时间内的信息，可以查看数据记录的时间戳，判断数据是否处于合理的时间范围内。

（六）可信性

评估数据的可信程度，即用户对数据的信任度。可以通过数据来源评估或专家验证等方法来提高数据的可信性。

（七）可访问性

体现数据的可获取性和易用性，即授权用户能否方便地访问和使用数据。

二、数据清洗的基本概念

数据清洗是数据预处理的关键环节，核心是对数据进行检查、清理和修正，以提升数据质量。数据清洗工作包括识别和处理缺失值、异常值、重复数据等，以及对数据进行转换、编码等操作，确保数据的准确性、完整性和一致性。

三、识别和处理缺失值

（一）识别缺失值

在 SPSS 中，可以通过查看数据的摘要统计信息（如个案数、缺失值数等），或者执行“Missing Value Analysis”命令来识别数据集中各变量的缺失值情况，确定缺失值的位置和数量。也可以通过数据视图，查看变量值为空或为特殊标记（如“.”“？”等）的数据行。

（二）处理缺失值

1. 删除含有缺失值的样本

如果缺失值较少，且删除后对分析结果没有显著影响，可以选择删除含有缺失值的样本。在 SPSS 中，该操作可通过执行“Select Cases”命令实现。

2. 填补均值

对于数值型变量，可通过执行“Compute Variable”命令，计算该变量的均值 [如 MEAN(年龄)] 填补缺失值。

3. 填补中位数

对于数值型变量，可通过执行“Compute Variable”命令，计算该变量的中位数 [如 MEDIAN (年龄)] 填补缺失值。

4. 填补众数

对于分类变量（如性别、饮食习惯等），可通过执行“Frequencies”命令查看其频率分布，确定众数并用以填补缺失值。

数据清洗 2

四、识别和处理异常值

（一）识别异常值

执行“Summaries of separate variables”命令，使用箱线图法识别异常值。观察箱线图中超出上下限的点，这些点即为异常值。还可以执行“Descriptive Statistics”→“Z−score”命令，计算每个数据点的 Z 分数（即该数据点与均值的差除以标准差），一般来说，Z 分数的绝对值大于 3 的数据点可视为异常值。

（二）处理异常值

（1）修正异常值。如果异常值由数据录入错误或测量误差等原因导致，可以尝试查找原始数据进行更正。例如，发现某位老年人的收缩压为 220，经过核实为数据录入错误，实际应为 120，可及时更正。

（2）剔除异常值。对于无法修正的异常值，如果这些异常值对分析结果影响较大，可将其剔除。在 SPSS 中，可以通过设置筛选条件，将异常值样本排除在分析之外。例如，剔除收缩压大于 180 的异常值，可执行“Select Cases”命令，将条件设为“收缩压 <=180”。

五、识别和处理重复数据

（一）识别重复数据

执行“Identify Duplicate Cases”命令识别重复数据。在弹出的对话框中，选择标识重复个案的

依据变量（通常为姓名、身份证号等多个关键变量），然后指定输出变量的名称和标签等信息。执行后，会在数据集中新增一个变量，用于标记重复的数据行。

（二）处理重复数据

对于重复数据，如果确定是完全重复的多余数据，可以直接删除。识别出重复数据后，可执行“Select Cases”命令，保留非重复个案并剔除重复个案。

六、数据转换

（一）对数转换

对于数值型变量，如果其分布呈现右偏态，可以通过对数转换使其更接近正态分布。在 SPSS 中，可执行“Compute Variable”命令，通过输入“LN(原变量名)”或“LG10(原变量名)”等对数表达式设置新变量。

（二）标准化转换

对于不同量纲的数值型变量，为消除量纲的影响，可以进行标准化转换。常用的方法有 Z 分数标准化。在 SPSS 中，可通过执行“Descriptive Statistics”→“Save standardized values as variables”命令，生成标准化后的变量。

七、数据编码

（一）分类变量的编码

对于分类变量（如性别、饮食习惯等），通常需要将其编码为数值型变量，以便进行后续的分析。例如，将性别编码为“0”（女性）、“1”（男性）；将饮食习惯编码为“1”（偏荤）、“2”（偏素）、“3”（均衡）等。在 SPSS 中，可以通过执行“Recode into Different Variables”或“Recode into Same Variables”命令实现分类变量的编码操作。

（二）日期型变量的编码

对于日期型变量（如数据收集时间），可以根据分析目的，将其转换为特定的数值型格式。例如，将日期转换为距离某个起始日期的天数等。在 SPSS 中，该转换可通过执行“Compute Variable”命令，并输入各类日期函数实现，从而完成日期型变量的编码和转换。

任务实施

数据清洗的 SPSS 软件实现 1

一、处理缺失值

（一）数值类型和日期类型的变量缺失处理

MISSING 函数可以用来检测数值类型和日期类型的变量是否缺失，如 MISSING(年龄)。

步骤 1 在 SPSS 中打开数据文件，执行“Transform”→“Compute Variable”命令，如图 2−1 所示。

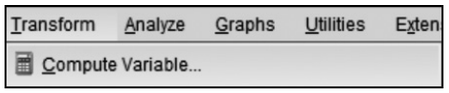
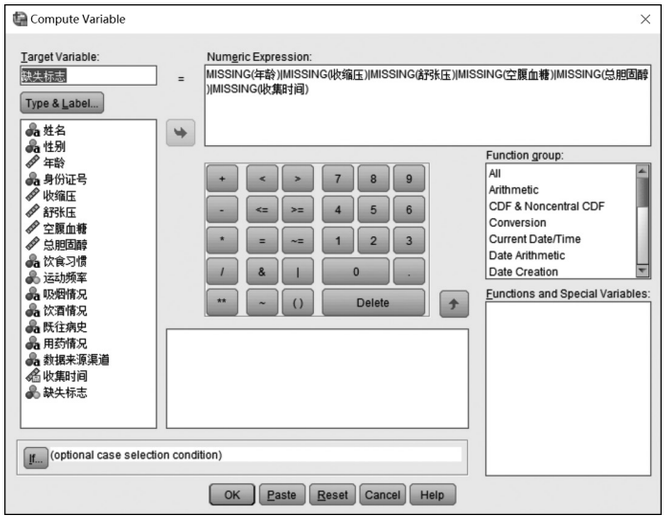


图 2-1 计算变量

步骤 2 在弹出的“Compute Variable”对话框的“Target Variable”文本框中输入新变量名“缺失标志”。

步骤 3 如果要同时查询多个数值类型或日期类型变量的缺失值，可以使用逻辑或运算符“|”连接各变量的缺失条件。在“Numeric Expression”文本框中输入缺失值判断公式“MISSING(年龄)|MISSING(收缩压)|MISSING(舒张压)|MISSING(空腹血糖)|MISSING(总胆固醇)|MISSING(收集时间)”，如图 2-2 所示。



数据清洗的 SPSS 软件实现 2

图 2-2 输入缺失值判断公式

步骤 4 单击“OK”按钮后，若“缺失标志”为 1，则表明该样本的“年龄、收缩压、舒张压、空腹血糖、总胆固醇、收集时间”等变量中至少存在一个缺失值；若“缺失标志”为 0，则代表不存在缺失值。经查询发现，第 21 条记录存在“年龄”缺失，第 47 条记录存在“收集时间”缺失，如图 2-3 所示。

	姓名	性别	年龄	身份证号	收缩压	舒张压	空腹血...	总胆固	饮食习惯	运动频	吸烟情	饮酒情况	既往病史	用药情况	数据来源渠道	收集时间	缺失标志
19		男	70	777777195501017757	136	83	5.80	4.60	均衡	4	否	否	高血压	缬沙坦	社区卫生...	2024/06/05	0
20	周电	女	76	777777194901017727	152	91	7.90	5.80	偏荤	1	是	是	糖尿病、高...	胰岛素...	医疗机构...	2024/06/13	0
21	吴明	男	.	777777196001017777	124	74	4.90	3.70	偏素	5	否	否	无	无	现场问卷...	2024/06/20	1

	姓名	性别	年龄	身份证号	收缩压	舒张压	空腹血...	总胆固	饮食习惯	运动频	吸烟情	饮酒情况	既往病史	用药情况	数据来源渠道	收集时间	缺失标志
46	陈王	女	70	777777195501017787	124	73	5.10	4.00	均衡	5	否	否	无	无	医疗机构...	2024/06/09	0
47	杨魏	男	77	777777194801017777	162	97	9.00	6.60	偏甜	0	是	是	糖尿病、高...	胰岛素...	小组访谈	.	1
48	赵放	女	68	777777195701017727	136	83	6.70	5.30	偏咸	1	是	是	糖尿病	二甲双...	社区卫生...	2024/06/14	0

图 2-3 查询出来的缺失值

步骤 5 因为只有个别样本的年龄数据缺失，可借助身份证号信息推算出年龄信息并予以补全。具体操作如下：执行“Transform”→“Compute Variable”命令，在打开的“Compute Variable”

对话框的“Target Variable”文本框中输入新变量名“推算年龄”。在“Numeric Expression”文本框中输入公式“2025−NUMBER(SUBSTR(身份证号 ,7,4),F8)”，该公式表示用当年年份减去身份证号中的出生年份。输入完成后，单击“OK”按钮，如图 2−4 所示。SPSS 会生成一个包含推算年龄的新变量“推算年龄”。

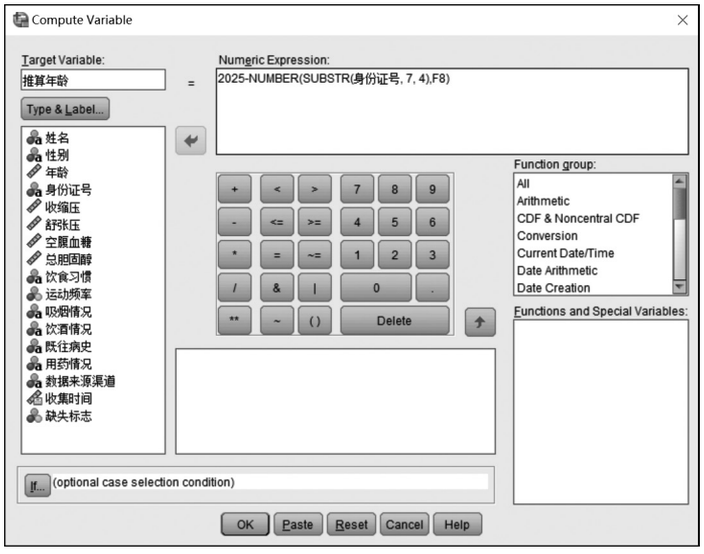


图 2−4　输入新变量名和公式 1

步骤 6　根据“推算年龄”列的值对“年龄”列缺失值进行填补即可，如图 2−5 所示。

姓名	性别	年龄	身份证号	收缩压	舒张压	空腹血...	总胆固...	饮食习...	运动频...	吸烟情	饮酒情	既往病...	用药情况	数据来源渠道	收集时间	缺失标志	推算年龄
陈图	女	130	777777195801017767	127	76	5.2	4.1	偏素	5	否	否	无	无	医疗机构...	2024/06/11	0	67
杨方	男	79	777777194601017777	165	98	9.2	6.8	偏甜	0	是	是	糖尿...	胰岛素、...	现场体检	2024/06/23	0	79
赵如	女	68	777777195701017767	142	84	6.8	5.4	偏咸	2	是	是	糖尿病	二甲双胍	小组访谈	2024/06/18	0	68
周电	女	76	777777194901017727	152	91	7.9	5.8	偏荤	1	是	是	糖尿...	胰岛素、...	医疗机构...	2024/06/13	0	76
吴明	男	65	777777196001017777	124	74	4.9	3.7	偏素	5	否	否	无	无	现场问卷...	2024/06/20	1	65

图 2−5　填补“年龄”列缺失值

步骤 7　“收集时间”缺失后，无法通过其他字段信息准确地填补，只能用数据采集周期的中位数日期进行填充。执行“Analyze”→“Descriptive Statistics”→“Frequencies”命令，系统弹出“Frequencies”对话框，将“收集时间”添加到“Variable(s)”框中，如图 2−6 和图 2−7 所示。

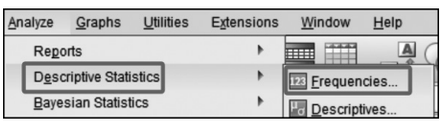


图 2−6　查看“收集时间”频率分布

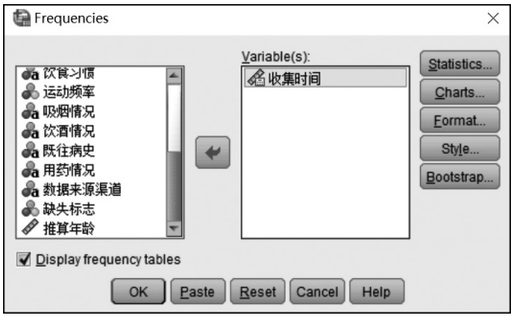


图 2−7　选择变量 1

步骤 8 在“Frequencies”对话框中，单击“Statistics”按钮，在弹出的“Frequencies: Statistics”对话框中选中“Median”复选框，然后单击“Continue”按钮，如图 2−8 所示。

步骤 9 单击“OK”按钮，SPSS 会输出“收集时间”数据的中位数，如图 2−9 所示。

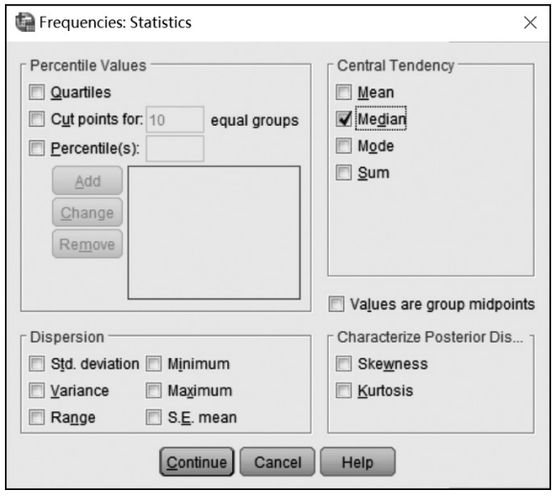
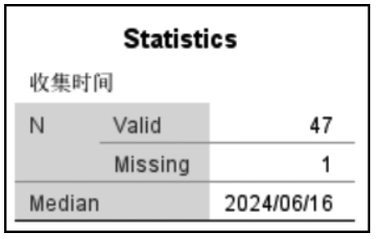


图 2−8 “Frequencies: Statistics”对话框 1



Statistics

收集时间

N	Valid	47
	Missing	1
Median		2024/06/16

图 2−9 输出“收集时间”数据的中位数

步骤 10 将中位数日期“2024/06/16”填充到第 47 条样本“收集时间”单元格中，如图 2−10 所示。

	情况	数据来源渠道	收集时间
40		社区卫生服务中心档案	2024/06/03
41		现场体检	2024/06/15
42	双胍	医疗机构检测报告	2024/06/28
43	阿司匹林	现场问卷调查	2024/06/30
44	地平	社区卫生服务中心档案	2024/06/12
45	坦	现场体检	2024/06/24
46		医疗机构检测报告	2024/06/09
47	平、阿司匹林	小组访谈	2024/06/16
48	阿卡波糖	社区卫生服务中心档案	2024/06/14

图 2−10 填充中位数日期

（二）字符串类型数据缺失值处理

MISSING 函数主要用于检测数值型变量的缺失值。对于字符串类型变量，该函数默认将空字符串视为非缺失值，因此无法准确识别字符串变量的缺失情况。

步骤 1 执行“Transform”→“Compute Variable”命令。

步骤 2 在弹出的“Compute Variable”对话框的“Target Variable”文本框中输入新变量名“缺失标志”。为同时检测字符串类型变量的缺失情况，可使用逻辑或运算符“|”将各变量的缺失条件连接起来。在“Numeric Expression”文本框中输入公式 " 姓名 ==""| 性别 ==""| 身份证号 ==""| 饮食习惯 ==""| 吸烟情况 ==""| 饮酒情况 ==""| 既往病史 ==""| 用药情况 ==""| 数据来源渠道 ==""，如图 2−11 所示。

图 2-11 输入新变量名和公式 2

步骤 3 单击“OK”按钮，该公式会对字符串类型数据是否为空字符串进行判断。若缺失标志为 1，则表示数据缺失；若缺失标志为 0，则表示数据不缺失。查询到有缺失值的样本如图 2-12 所示。

	姓名	性别	年龄	身份证号	收缩压	舒张压	空腹血...	总胆固	饮食习惯	运动频	吸烟情	饮酒情况	既往病史	用药情况	数据来源渠道	收集时间	缺失标志
10	钱封	女	-10	777777195701017747	128	75	5.40	4.30	偏素	3	否	否	无	无	社区卫生...	2024/05/25	0
11	孙德	男	69	777777195601017777	138	86	6.50	5.20	偏荤	1	是	是	高血压	美托洛尔	现场体检	2024/06/19	0
12	李为	女	74	777777195101017707	148	89	7.70	5.70	偏甜	0	否	否	糖尿病	二甲双...	医疗机构...	2024/06/14	0
13	王齐		66	777777195901017747	122	72	5.00	3.90	均衡	4	否	否	无	无	小组访谈	2024/06/21	1
14	张杨	女	77	777777194801017727	158	10	8.40	6.20	偏咸	1	是	是	糖尿病、冠...	胰岛素...	现场问卷...	2024/06/16	0
15	刘科	男	73	777777195201017777	132	81	5.60	4.50	偏荤	3	是	否	高血压	氯氯地平	社区卫生...	2024/05/30	0
16	陈图	女	130	777777195801017767	127	76	5.20	4.10	偏素	5	否	否	无	无	医疗机构...	2024/06/11	0
17	杨方	男	79	777777194601017777	165	98	9.20	6.80	偏甜	0	是	是	糖尿病、高...	胰岛素...	现场体检	2024/06/23	0
18	赵如	女	68	777777195701017767	142	84	6.80	5.40	偏咸	2	是	是	糖尿病	二甲双胍	小组访谈	2024/06/18	0
19		男	70	777777195501017757	136	83	5.80	4.60	均衡	4	否	否	高血压	缬沙坦	社区卫生...	2024/06/05	1
20	周电	女	76	777777194901017727	152	91	7.90	5.80	偏荤	1	是	是	糖尿病、高...	胰岛素...	医疗机构...	2024/06/13	0

图 2-12 查询到有缺失值的样本

步骤 4 若“姓名”缺失且无法填充，可直接删除该样本。若“性别”缺失，可从“身份证号”中获取相关信息：身份证号第 17 位若为偶数，代表该样本为女性；若为奇数，代表该样本为男性。

步骤 5 执行“Transform”→“Compute Variable”命令，在弹出的“Compute Variable”对话框的“Target Variable”文本框中输入新变量名“性别数字”。接着在“Numeric Expression”文本框中输入公式“NUMBER(SUBSTR(身份证号 ,17,1),F8)”，如图 2-13 所示。

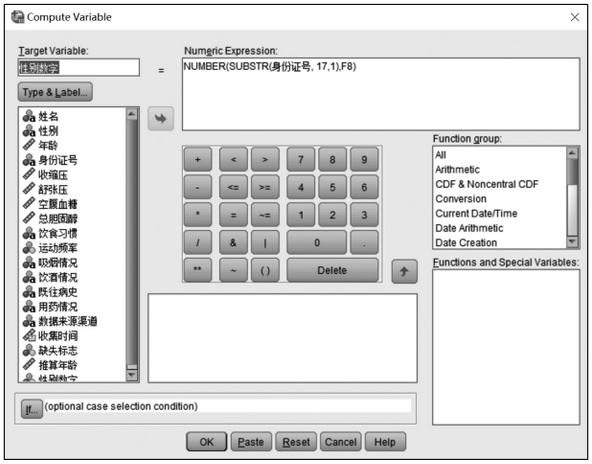


图 2-13 输入新变量名和公式 3

步骤 6 单击“OK”按钮，完成“性别数字”的提取。提取的数字为偶数代表女性，为奇数则代表男性，由此完成性别缺失值的填补，如图 2-14 所示。

姓名	性别	年龄	身份证号	收缩压	舒张压	空腹血...	总胆固...	饮食习...	运动频...	吸烟情	饮酒情	既往病...	用药情况	数据来源渠道	收集时间	缺失标志	推算年龄	缺失标志2	性别数字
周九	男	63	777777196201017777	125	78	5.1	4.0	均衡	5	否	否	无	无	医疗机构...	2024/06/12	0	63	0	7
赵放	女	68	777777195701017727	136	83	6.7	5.3	偏咸	1	是	是	糖尿病	二甲双胍...	社区卫生...	2024/06/14	0	68	0	2
郑红	男	76	777777194901017757	190	92	8.1	6.0	偏咸	2	是	否	糖尿...	胰岛素、...	现场问卷...	2024/06/22	0	76	0	5
钱封	女	-10	777777195701017747	128	75	5.4	4.3	偏素	3	否	否	无	无	社区卫生...	2024/05/25	0	68	0	4
孙德	男	69	777777195601017777	138	86	6.5	5.2	偏荤	1	是	是	高血压	美托洛尔	现场体检	2024/06/19	0	69	0	7
李为	女	74	777777195101017707	148	89	7.7	5.7	偏甜	0	否	否	糖尿病	二甲双胍...	医疗机构...	2024/06/14	0	74	0	0
王齐	女	66	777777195901017747	122	72	5.0	3.9	均衡	4	否	否	无	无	小组访谈	2024/06/21	0	66	1	4

图 2-14 通过“性别数字”确定性别

二、识别异常值

（一）数字类型数据

数字类型数据的异常值检测是一个多方法体系，描述统计法、箱线图法和 Z 分数法是三种常见的方法，但它们的原理和逻辑是不同的。描述统计法主要基于数据的集中趋势和离散程度，箱线图法基于四分位数和四分位距（IQR），而 Z 分数法则基于数据的标准化处理。尽管三种方法的目标都是识别异常值，但由于方法不同，结果可能不一致，需根据实际规定的异常范围进一步核对确认。

1. 描述统计法

步骤 1 执行“Analyze”→“Descriptive Statistics”→“Descriptives”命令，如图 2-15 所示。

图 2-15 执行“Analyze”→“Descriptive Statistics”→“Descriptives”命令 1

步骤 2 在弹出的“Descriptives”对话框中，将“年龄”“空腹血糖”等数字变量选入右侧的“Variable(s)”框中，如图 2-16 所示。

步骤 3 单击“Options”按钮，系统弹出“Descriptives: Options”对话框，选中“Mean”“Std. deviation”“Minimum”“Maximum”“Skewness”“Kurtosis”等统计量复选框，然后单击“Continue”按钮，如图 2-17 所示。然后，系统会返回“Descriptives”对话框，再单击“OK”按钮。

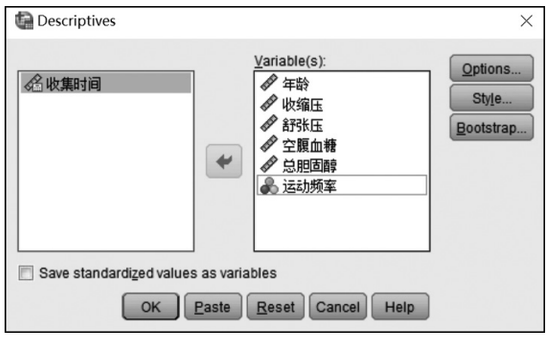


图 2-16 选择变量 2

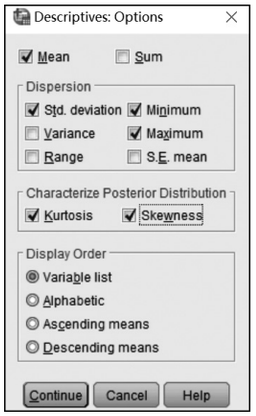


图 2-17 选中所需的统计量复选框

步骤 4 查看输出结果中的最小值和最大值，可初步判断是否存在超出合理范围的异常值。具体判定标准如下：若年龄超出 0 ～ 120 岁范围，收缩压超出 60 ～ 200 mmHg 范围，舒张压超出 40 ～ 120 mmHg 范围，血糖超出 2.2 ～ 22 mmol/L 范围，总胆固醇超出 2 ～ 10 mmol/L 范围，运动频率超出 0 ～ 14 次，均判定为异常值。输出结果中的异常值情况如图 2−18 所示。

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
年龄	48	-10	130	71.15	15.188	-2.088	.343	21.959	.674
收缩压	48	120	220	143.77	18.051	1.902	.343	6.024	.674
舒张压	48	10	100	83.21	13.042	-3.802	.343	21.092	.674
空腹血糖	48	4.5	28.0	7.260	3.7114	4.351	.343	22.118	.674
总胆固醇	48	1.0	12.0	5.110	1.4606	1.835	.343	10.556	.674
运动频率	48	0	5	2.15	1.688	.342	.343	-1.159	.674
Valid N (listwise)	48								

图 2−18 输出结果中的异常值情况

2. 箱线图法

上述异常值还可以通过箱线图法进一步确定。

步骤 1 执行“Graphs”→“Legacy Dialogs”→“Boxplot”命令，如图 2−19 所示。

步骤 2 系统弹出“Boxplot”对话框，单击“Simple”按钮，选中“Summaries of separate variables”单选按钮，单击“Define”按钮，如图 2−20 所示。

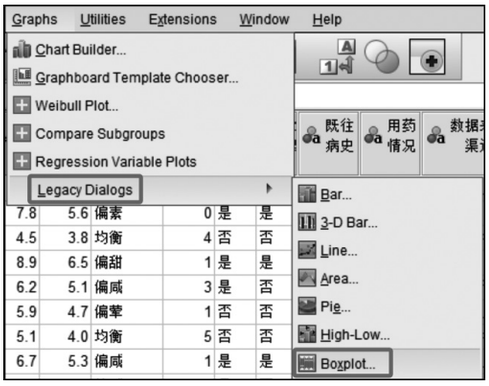


图 2−19 制作箱线图

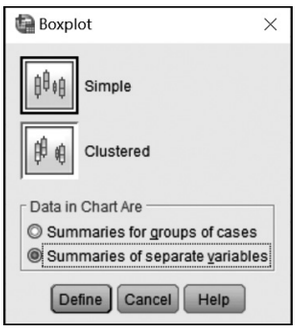


图 2−20 “Boxplot”对话框

步骤 3 在弹出的“Define Simple Boxplot:Summaries of Separate Variables”对话框中，将“年龄”“空腹血糖”等数字变量移入“Boxes Represent”框中，然后单击“OK”按钮，如图 2−21 所示。

步骤 4 箱线图（图 2−22）会显示数据的中位数、四分位数和异常值，其中超出箱线图上下边缘的点被判定为异常值。从图 2−22 中可知，第 10 个、第 16 个样本的年龄，第 9 个、第 23 个样本的收缩压，第 14 个样本的舒张压，第 35 个、第 43 个样本的空腹血糖，第 41 个、第 38 个样本的总胆固醇均被初步认定为异常值，不过这些结果是否准确，还需进一步确认。

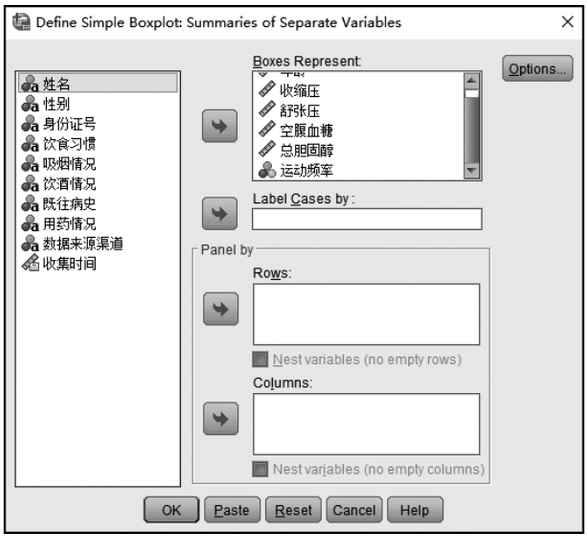


图 2-21 选择变量 3

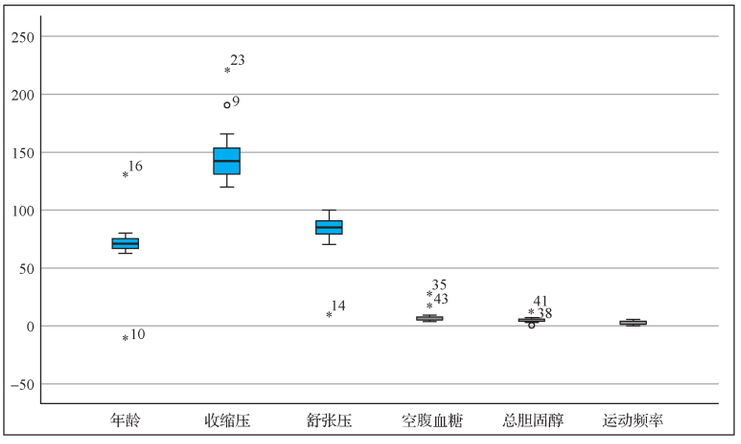


图 2-22 生成的箱线图

3. Z 分数法

步骤 1 执行“Analyze”→“Descriptive Statistics”→“Descriptives”命令。

步骤 2 在弹出的“Descriptives”对话框中，将“年龄”“空腹血糖”等数字变量移入右侧的“Variable(s)”框中，然后选中“Save standardized values as variables”复选框，如图 2-23 所示。

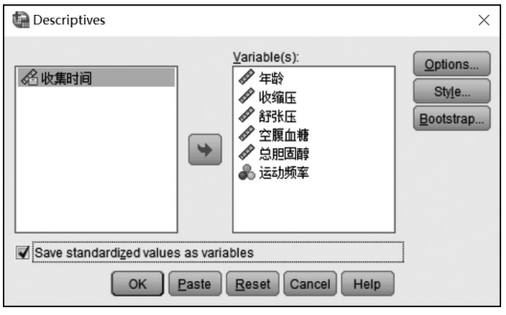


图 2-23 选择变量 4

步骤 3 单击“OK”按钮，SPSS 会在数据视图中生成对应的 Z 分数变量（如“Z 年龄”“Z 空腹血糖”“Z 舒张压”），如图 2-24 所示。

	用药情况	数据来源渠道	收集时间	Z年龄	Z收缩压	Z空腹血糖	Z舒张压	Z总胆固醇	Z运动频率
1	氯地平	社区卫生...	2024/06/10	-.20651	-.49074	-.53175	.13558	-.62484	-.06317
2	二甲双...	现场体检	2024/06/15	.05405	.33301	.13578	.51486	.32469	-1.25073
3	无	医疗机构...	2024/05/20	.24947	-1.31449	-.74535	-1.00229	-.89614	1.12440
4	胰岛素...	小组访谈	2024/06/20	-.40193	.88217	.42949	1.27344	.93510	-.65695
5	他托洛尔	现场问卷...	2024/06/18	-.07623	-.21616	-.29144	-.24371	-.01443	.53061

图 2-24 Z 分数变量

步骤 4 一般情况下，Z 分数绝对值大于 3 可判定为异常值。通过筛选 Z 分数大于 3 或小于 -3 的记录，可以找到异常值。执行“Data”→“Select Cases”命令，如图 2-25 所示。

步骤 5 在弹出的“Select Cases”对话框中选中“If condition is satisfied”单选按钮，再单击“If”按钮，如图 2-26 所示。在打开的“Select Cases: If”对话框中，输入公式“ABS(Z 年龄)>3|ABS(Z 收缩压)>3|ABS(Z 空腹血糖)>3|ABS(Z 舒张压)>3|ABS(Z 总胆固醇)>3|ABS(Z 运动频率)>3”，如图 2-27 所示。

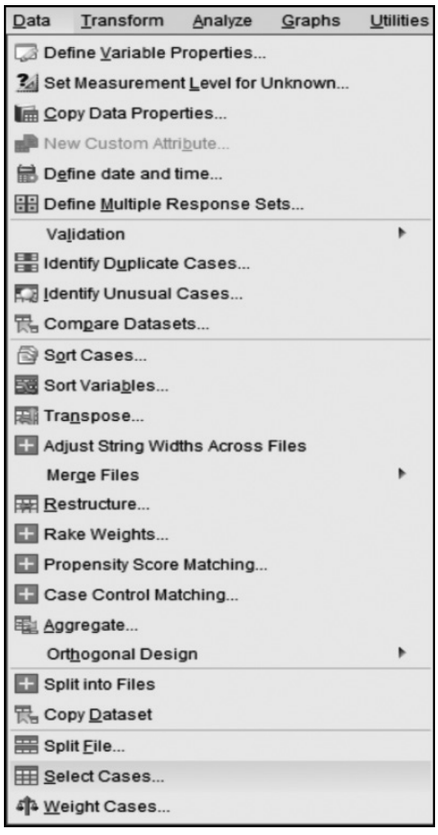


图 2-25 选择个案

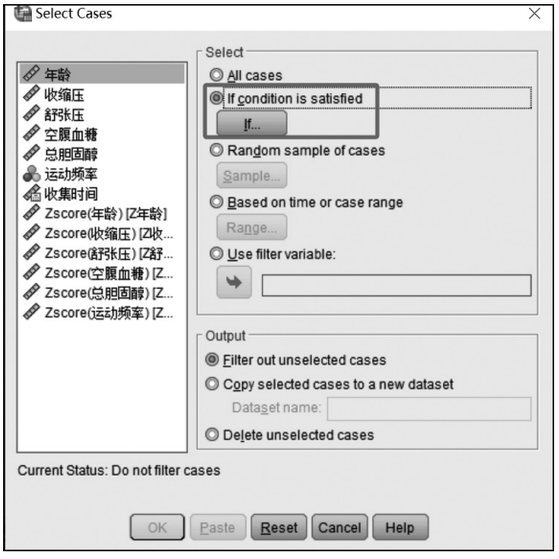


图 2-26 “Select Cases”对话框 1

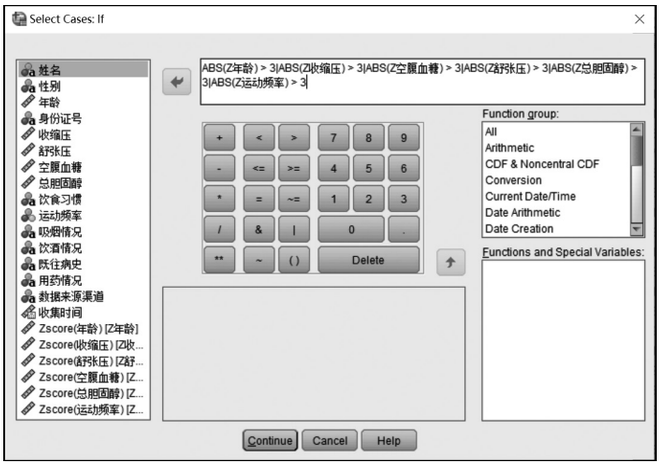


图 2-27 输入公式

步骤 6 生成新变量，新变量值为 1 的样本代表存在异常，如图 2-28 所示。

	Z年龄	Z收缩压	Z空腹血糖	Z舒张压	Z总胆固醇	Z运动频率	filter_$
10	-5.28747	-.87516	-.50504	-.62300	-.55702	.53061	1
11	-.14137	-.32599	-.21133	.21143	.05339	-.65695	0
12	.18433	.22317	.10908	.43901	.39251	-1.25073	0
13	-.33679	-1.20466	-.61185	-.85058	-.82832	1.12440	0
14	.37976	.77234	.29598	-5.55376	.73163	-.65695	1
15	.11919	-.65549	-.45164	-.16786	-.42137	.53061	0
16	3.83221	-.93008	-.55845	-.54714	-.69267	1.71818	1
17	.51004	1.15675	.50959	1.12173	1.13857	-1.25073	0
18	-.20651	-.10633	-.13123	.05972	.18904	-.06317	0
19	.31462	.44284	.16248	.59072	.46034	-.65695	0
20	-.40193	-1.09483	-.63855	-.69886	-.96396	1.71818	0
21	.05405	.00351	-.23804	.28729	-.08225	-1.25073	0
22	.37976	.66251	.26928	.74244	.66381	-.65695	0
23	-.14137	4.17717	-.47834	-.24371	-.48920	.53061	1

	Z年龄	Z收缩压	Z空腹血糖	Z舒张压	Z总胆固醇	Z运动频率	filter_$
34	-.07623	.00351	-.07783	.28729	.05339	-.06317	0
35	.37976	.88217	5.52936	.97001	.79945	-1.25073	1
36	-.14137	-.76533	-.50504	-.09200	-.55702	1.12440	0
37	.18433	.44284	.05567	.59072	.32469	-.65695	0
38	-.01109	.22317	-.10453	.36315	-2.79520	.53061	0
39	.31462	-.54566	-.55845	-.39543	-.69267	1.71818	0
40	-.33679	-.87516	-.66525	-.54714	-1.03179	1.12440	0
41	.05405	.11334	-.05113	.43901	4.66541	-1.25073	1

图 2-28 异常样本

步骤 7 由图 2-28 可知：第 10 个、第 16 个样本的年龄，第 14 个样本的舒张压，第 23 个样本的收缩压，第 35 个样本的空腹血糖，第 41 个样本的总胆固醇，均被认定为异常值。

步骤 8 对于异常值，可进行修改或删除操作。例如，“年龄”可根据身份证号确定，“血压”“血糖”等值可通过与本人沟通进行核对。在此，将第 10 个、第 16 个样本的年龄分别修改为 68 岁、67 岁，将第 23 个样本的收缩压改为 200 mmHg，将第 14 个样本的舒张压改为 40 mmHg，将第 35 个样本的空腹血糖改为 22 mmol/L，将第 41 个样本的总胆固醇改为 10 mmol/L。

（二）字符串类型数据

1. 频率分析

步骤 1 执行“Analyze”→“Descriptive Statistics”→“Frequencies”命令，如图 2-29 所示。

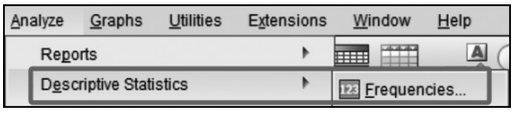


图 2-29 分析数据频率

步骤 2 在弹出的“Frequencies”对话框中将“姓名”“性别”等字符串类型变量移入“Variable(s)”框中，选中“Display frequency tables”复选框，单击“OK”按钮，如图 2-30 所示。

图 2-30 选择字符串类型变量

步骤 3 查看输出结果中的频率表，如图 2-31 至图 2-38 所示。检查是否存在不符合预期的类别。例如，性别通常只有“男”和“女”两个类别，如果出现其他类别（如“未知”“未填写”等），则可能是异常值。在这里性别存在异常情况，需要查找属于异常类别（如“未知”“未填写”等）的是哪一个样本记录。

性别

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	1	2.1	2.1	2.1
	男	21	44.7	44.7	46.8
	男性	1	2.1	2.1	48.9
	女	24	51.1	51.1	100.0
	Total	47	100.0	100.0	

图 2-31 “性别”频率表

饮食习惯

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	均衡	9	19.1	19.1	19.1
	偏荤	10	21.3	21.3	40.4
	偏素	8	17.0	17.0	57.4
	偏甜	9	19.1	19.1	76.6
	偏咸	11	23.4	23.4	100.0
	Total	47	100.0	100.0	

图 2-32 “饮食习惯”频率表

姓名

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	陈花	1	2.1	2.1	2.1
	陈金	1	2.1	2.1	4.3
	陈图	1	2.1	2.1	6.4
	陈王	1	2.1	2.1	8.5
	黄刚	1	2.1	2.1	10.6
	黄秋	1	2.1	2.1	12.8
	李四	1	2.1	2.1	14.9
	李为	1	2.1	2.1	17.0
	刘春	1	2.1	2.1	19.1
	刘科	2	4.3	4.3	23.4
	刘烈	1	2.1	2.1	25.5
	刘七	1	2.1	2.1	27.7
	钱封	1	2.1	2.1	29.8
	孙八	1	2.1	2.1	31.9
	孙德	1	2.1	2.1	34.0
	王平	1	2.1	2.1	36.2
	王齐	1	2.1	2.1	38.3
	王全	1	2.1	2.1	40.4
	王五	1	2.1	2.1	42.6
	王雨	1	2.1	2.1	44.7
	吴乐	1	2.1	2.1	46.8
	吴路	1	2.1	2.1	48.9
	吴明	1	2.1	2.1	51.1
	杨方	1	2.1	2.1	53.2
	杨丘	1	2.1	2.1	55.3
	杨魏	1	2.1	2.1	57.4
	杨越	1	2.1	2.1	59.6
	张国	1	2.1	2.1	61.7
	张政	1	2.1	2.1	63.8
	张三	1	2.1	2.1	66.0
	张想	1	2.1	2.1	68.1
	张杨	1	2.1	2.1	70.2
	赵东	1	2.1	2.1	72.3
	赵放	2	4.3	4.3	76.6
	赵丰	1	2.1	2.1	78.7
	赵六	1	2.1	2.1	80.9
	赵如	1	2.1	2.1	83.0
	郑红	1	2.1	2.1	85.1
	郑秋	1	2.1	2.1	87.2
	郑新	1	2.1	2.1	89.4
	郑阳	1	2.1	2.1	91.5
	周电	1	2.1	2.1	93.6
	周九	1	2.1	2.1	95.7
	周山	1	2.1	2.1	97.9
	周新	1	2.1	2.1	100.0
	Total	47	100.0	100.0	

身份证号

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	777777194601017777	1	2.1	2.1	2.1
	777777194701017777	1	2.1	2.1	4.3
	777777194701017787	1	2.1	2.1	6.4
	777777194801017727	1	2.1	2.1	8.5
	777777194801017757	1	2.1	2.1	10.6
	777777194801017777	1	2.1	2.1	12.8
	777777194801017787	1	2.1	2.1	14.9
	777777194901017727	1	2.1	2.1	17.0
	777777194901017757	1	2.1	2.1	19.1
	777777194901017777	1	2.1	2.1	21.3
	777777194901017787	1	2.1	2.1	23.4
	777777195001017757	1	2.1	2.1	25.5
	777777195001017777	1	2.1	2.1	27.7
	777777195001017787	1	2.1	2.1	29.8
	777777195101017707	1	2.1	2.1	31.9
	777777195101017777	1	2.1	2.1	34.0
	777777195101017787	1	2.1	2.1	36.2
	777777195201017757	1	2.1	2.1	38.3
	777777195201017777	1	2.1	2.1	40.4
	777777195201017787	1	2.1	2.1	42.6
	777777195301017727	1	2.1	2.1	44.7
	777777195301017747	1	2.1	2.1	46.8
	777777195301017787	1	2.1	2.1	48.9
	777777195301017797	1	2.1	2.1	51.1
	777777195401017777	1	2.1	2.1	53.2
	777777195401017787	1	2.1	2.1	55.3
	777777195501017737	1	2.1	2.1	57.4
	777777195501017777	1	2.1	2.1	59.6
	777777195501017787	1	2.1	2.1	61.7
	777777195601017757	1	2.1	2.1	63.8
	777777195601017777	1	2.1	2.1	66.0
	777777195601017787	1	2.1	2.1	68.1
	777777195701017727	2	4.3	4.3	72.3
	777777195701017747	1	2.1	2.1	74.5
	777777195701017767	1	2.1	2.1	76.6
	777777195701017777	1	2.1	2.1	78.7
	777777195701017787	1	2.1	2.1	80.9
	777777195801017767	1	2.1	2.1	83.0
	777777195801017777	1	2.1	2.1	85.1
	777777195801017787	1	2.1	2.1	87.2
	777777195901017727	1	2.1	2.1	89.4
	777777195901017747	1	2.1	2.1	91.5
	777777195901017777	1	2.1	2.1	93.6
	777777196001017777	1	2.1	2.1	95.7
	777777196001017787	1	2.1	2.1	97.9
	777777196201017777	1	2.1	2.1	100.0
	Total	47	100.0	100.0	

图 2-33 “姓名”和“身份证号”频率表

吸烟情况

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	否	16	34.0	34.0	34.0
	是	31	66.0	66.0	100.0
	Total	47	100.0	100.0	

图 2-34 “吸烟情况”频率表

饮酒情况

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	否	25	53.2	53.2	53.2
	是	22	46.8	46.8	100.0
	Total	47	100.0	100.0	

图 2-35 “饮酒情况”频率表

数据来源渠道

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	社区卫生服务中心档案	13	27.7	27.7	27.7
	现场体检	8	17.0	17.0	44.7
	现场问卷调查	7	14.9	14.9	59.6
	小组访谈	7	14.9	14.9	74.5
	医疗机构检测报告	12	25.5	25.5	100.0
	Total	47	100.0	100.0	

图 2-36 “数据来源渠道”频率表

既往病史

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	高血压	1	2.1	2.1	2.1
	高血压	10	21.3	21.3	23.4
	高血压症	1	2.1	2.1	25.5
	糖尿病	10	21.3	21.3	46.8
	糖尿病、高血压	5	10.6	10.6	57.4
	糖尿病、高血压、冠心病	2	4.3	4.3	61.7
	糖尿病、高血压、脑梗死	2	4.3	4.3	66.0
	糖尿病、冠心病	5	10.6	10.6	76.6
	无	11	23.4	23.4	100.0
	Total	47	100.0	100.0	

图 2-37 “既往病史”频率表

用药情况

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	氨氯地平	4	8.5	8.5	8.5
	二甲双胍	4	8.5	8.5	17.0
	二甲双胍、阿卡波糖	4	8.5	8.5	25.5
	二甲双胍、阿司匹林	1	2.1	2.1	27.7
	二甲双胍、格列齐特	2	4.3	4.3	31.9
	美托洛尔	4	8.5	8.5	40.4
	无	11	23.4	23.4	63.8
	缬沙坦	4	8.5	8.5	72.3
	胰岛素、阿司匹林	4	8.5	8.5	80.9
	胰岛素、安氯地平、阿司匹林	1	2.1	2.1	83.0
	胰岛素、氨氯地平	2	4.3	4.3	87.2
	胰岛素、氨氯地平、阿司匹林	2	4.3	4.3	91.5
	胰岛素、缬沙坦	3	6.4	6.4	97.9
	胰岛素、缬沙坦、阿司匹林	1	2.1	2.1	100.0
	Total	47	100.0	100.0	

图 2-38 “用药情况”频率表

2. 数据清理

步骤 1 如果发现性别变量中存在异常值，可执行“Data”→“Select Cases”命令，在弹出的“Select Cases”对话框中选中“If condition is satisfied”单选按钮，如图 2-39 所示。

步骤 2 在弹出的“Select Cases: If”对话框中，输入筛选条件“性别 =' 男性 '”，如图 2-40 所示。然后单击“Continue”→“OK”按钮。筛选后的结果如图 2-41 所示。

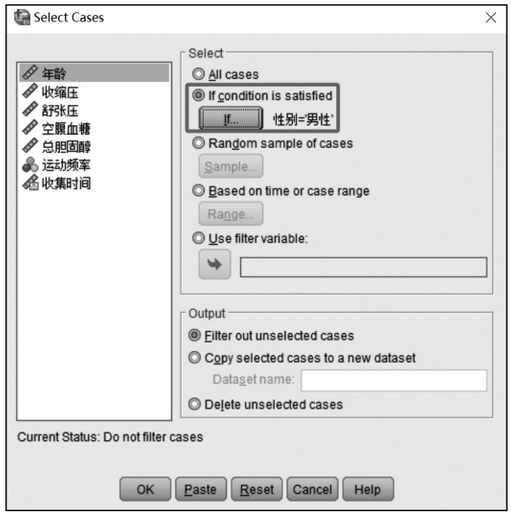


图 2-39 “Select Cases”对话框 2

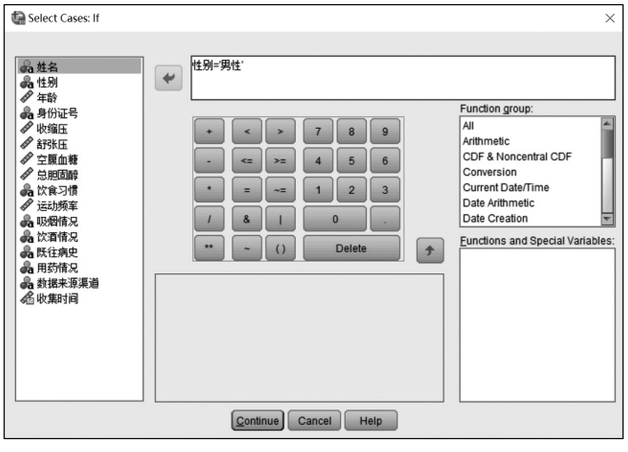


图 2-40 输入筛选条件

	姓名	性别	年龄	身份证号	收缩压	舒张压	空腹血糖	总胆固	饮食习惯	运动频	吸烟情	饮酒情况	既往病史	用药情况	数据来源渠道	收集时间	filter_$
24	烈	男	74	777777195101017777	146	88	7.20	5.50	偏荤	2	是	是	糖尿病	二甲双...	小组访谈	2024/06/22	0
25	花	女	71	777777195401017787	129	77	5.30	4.20	偏素	4	否	否	高血压	缬沙坦	社区卫生...	2024/06/08	0
26	越	男	78	777777194701017777	162	95	8.80	6.40	偏甜	0	是	是	糖尿病、高...	胰岛素...	现场体检	2024/06/27	0
27	丰	女	66	777777195901017727	134	82	6.10	4.80	偏咸	1	是	是	糖尿病	二甲双胍	医疗机构...	2024/06/19	0
28	刚	男性	72	777777195301017797	140	85	5.70	4.50	均衡	3	是	否	高血压症	氨氯地平	现场问卷...	2024/06/14	1

图 2-41 筛选后的结果

步骤 3 对筛选出的异常值记录进行处理，如进行删除操作或重新赋值。可根据身份证号的第 17 位数字的奇偶性来确定性别，这里将“男性”改为“男”。

（三）日期类型数据

日期类型数据的异常值可以通过格式检查、逻辑检查两种方式来识别。

1. 格式检查

步骤 1 确保“收集时间”变量的格式正确。如果日期格式不正确，可单击“Variable View”，找到需要修改的变量所在的行，在“Type”列中，单击“Date”单元格右侧的按钮，如图 2-42 所示。

步骤 2 在弹出的“Variable Type”对话框中，选中“Date”单选按钮，如图 2-43 所示。如果日期格式与默认格式不符，则需要进一步调整日期格式。

图 2-42 单击“Date”单元格右侧的按钮

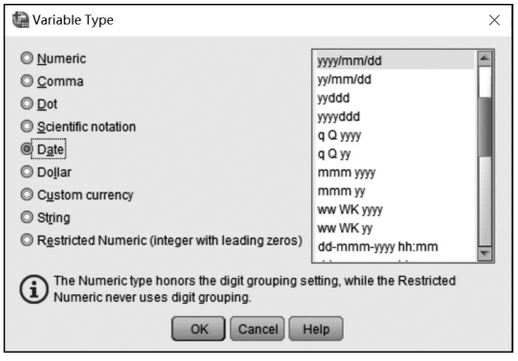


图 2-43 “Variable Type”对话框

2. 逻辑检查

步骤 1 执行“Analyze”→“Descriptive Statistics”→“Descriptives”命令，打开“Descriptives”对话框，将“收集时间”变量选入“Variable(s)”框中，如图 2-44 所示。单击“Options”按钮，系统弹出“Descriptives: Options”对话框，选中“Minimum”“Maximum”和“Mean”等复选框，再单击“Continue”→“OK”按钮。即可查看“收集时间”的各项值，如图 2-45 所示。

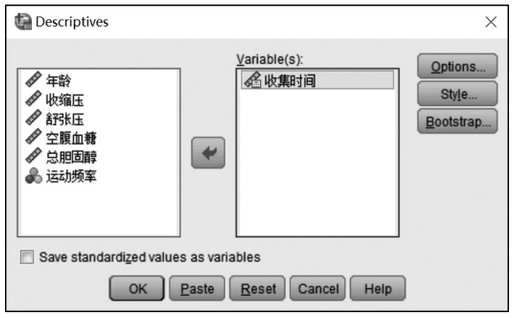


图 2-44 选择变量 5

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
收集时间	47	2024/05/15	2026/06/23	2024/06/29	108 06:30:5...	6.754	.347	46.065	.681
Valid N (listwise)	47								

图 2-45 查看“收集时间”的各项值

步骤 2 从图 2−45 中可以看出“收集时间”最大值是 2026/06/23，最小值是 2024/05/15，均值是 2024/06/29。在很多研究项目中，数据收集时间是预先规划好的，可以通过“收集时间”的均值、最大值和最小值来检查是否存在明显不合逻辑的日期，例如，未来的日期或远早于研究开始时间的日期。在这里，2026/06/23 就不符合规律。

步骤 3 如果发现异常日期，可以执行“Data”→“Select Case”命令，在弹出的“Select Cases”对话框中选中“If condition is satisfied”单选按钮，再单击“If”按钮，如图 2−46 所示。

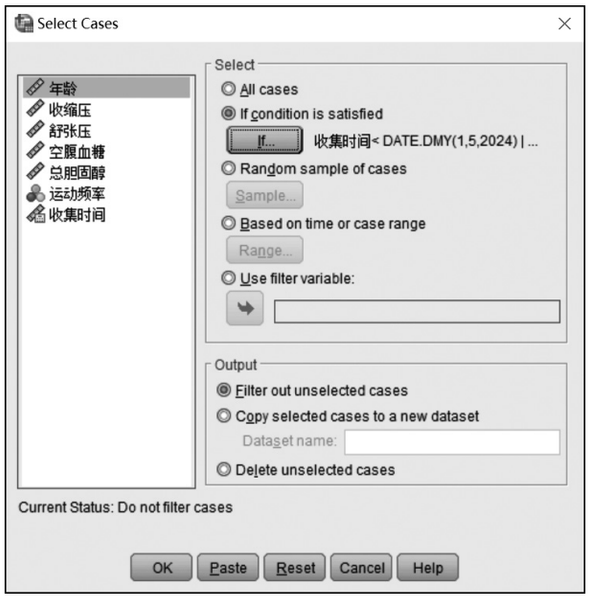


图 2−46 “Select Cases”对话框 3

步骤 4 系统弹出“Select Cases: If”对话框，在其文本框中输入公式“收集时间 <DATE.DMY(1,5,2024)| 收集时间 >DATE.DMY(1,7,2024)”，将收集日期设置为 2024 年 5 月 1 日至 2024 年 7 月 1 日，如图 2−47 所示。

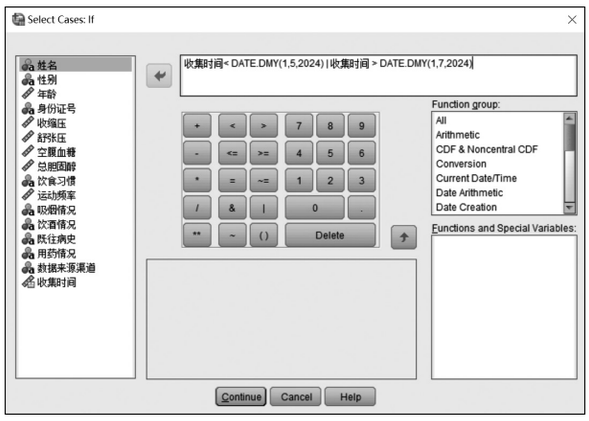


图 2−47 输入公式 2

步骤 5 若日期超出“2024.5.1—2024.7.1”范围，则判定为异常值。从生成的结果变量取值情况可以识别出异常值，当该结果变量取值为 1 时，表明对应记录的日期为异常值；当取值为 0 时，

表明对应记录的日期为正常值。对于异常值，可通过核查进行修改，或者直接删除对应记录。“收集时间”的异常值如图 2−48 所示。

	空腹血糖	总胆固	饮食习惯	运动频	吸烟情	饮酒情况	既往病史	用药情况	数据来源渠道	收集时间	filter_$
31	6.70	5.30	偏甜	0	是	是	糖尿病	二甲双...	小组访谈	2024/06/24	0
32	8.50	6.20	偏咸	1	是	否	糖尿病、冠...	胰岛素...	社区卫生...	2024/06/29	0
33	5.60	4.40	均衡	3	否	否	高血压	缬沙坦	现场体检	2024/06/21	0
34	7.00	5.20	偏荤	2	是	是	糖尿病	二甲双胍	医疗机构...	2024/06/10	0
35	28.00	6.30	偏甜	0	是	是	糖尿病、高...	胰岛素...	现场问卷...	2024/06/26	0
36	5.40	4.30	偏咸	4	是	否	高血压	美托洛尔	社区卫生...	2024/06/07	0
37	7.50	5.60	均衡	1	是	是	糖尿病、高...	胰岛素...	医疗机构...	2024/06/18	0
38	6.90	1.00	偏荤	3	是	是	糖尿病	二甲双...	小组访谈	2024/06/23	0
39	5.20	4.10	偏素	5	否	否	无	无	社区卫生...	2026/06/03	1
40	4.80	3.60	均衡	4	否	否	无	无	现场体检	2024/06/15	0

图 2−48 “收集时间”的异常值

三、识别重复数据

步骤 1 执行“Data”→“Identify Duplicate Cases”命令，如图 2−49 所示。

步骤 2 在弹出的“Identify Duplicate Cases”对话框中，将“姓名”“性别”“身份证号”等用于标识重复个案的关键变量移入“Define matching cases by”框中。选中“Indicator of primary cases(1=unique or primary,0=duplicate)”“Move matching cases to the top of the file”“Display frequencies for created variables”复选框，然后设置输出变量的名称为“重复个案”，如图 2−50 所示。

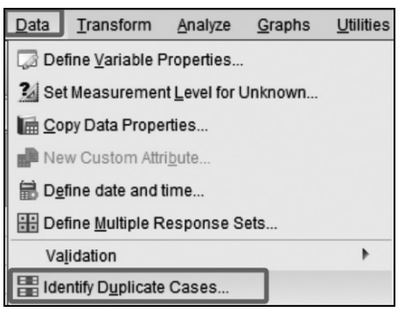


图 2−49 执行“Data”→“Identify Duplicate Cases”命令

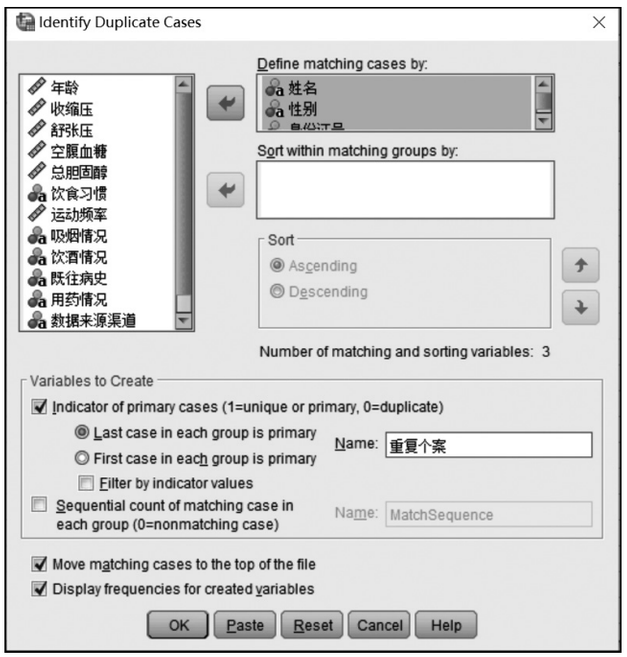


图 2−50 “Identify Duplicate Cases”对话框

步骤 3 执行上述操作后，数据集将新增一个变量，用于标记重复的数据行。如果确定是完全重复的多余数据，可将其直接删除，如图 2−51 所示。

	姓名	性别	年龄	身份证号	收缩压	舒张压	空腹血糖	总胆固	饮食习	运动频	吸烟情	饮酒情况	既往病史	用药情况	数据来源渠道	收集时间	重复个案
1	赵放	女	68	777777...	136	83	6.70	5.30	偏咸	1	是	是	糖尿病	二甲双...	社区卫生...	2024/06/14	0
2	赵放	女	68	777777...	136	83	6.70	5.30	偏咸	1	是	是	糖尿病	二甲双...	社区卫生...	2024/06/14	1
3	陈花	女	71	777777...	129	77	5.30	4.20	偏素	4	否	否	高血压	缬沙坦	社区卫生...	2024/06/08	1
4	陈金	女	77	777777...	160	96	28.00	6.30	偏甜	0	是	是	糖尿病、高...	胰岛素...	现场问卷...	2024/06/26	1

图 2-51 重复数据行

知识拓展

一、数据清洗的重要性

健康数据的准确性、完整性和一致性，对医疗决策、临床研究、公共卫生政策制定等场景至关重要。数据清洗是确保健康数据质量的关键步骤，能够帮助用户剔除错误数据、不完整数据及不一致的数据，为后续的数据分析和应用提供可靠基础。

二、健康数据清洗面临的挑战

（一）数据的异构性和多源性

健康数据来源多样，包括电子病历系统、医疗物联网设备、实验室信息系统、可穿戴设备等。这些数据在结构、格式、语义等方面存在显著差异，增加了数据清洗与融合的难度。

（二）数据量大且复杂

随着医疗信息化的不断发展，健康数据量呈爆炸式增长，且数据类型复杂多样，包括结构化数据、半结构化数据和非结构化数据。对如此庞大和复杂的健康数据进行清洗，需要高效的清洗算法和强大的计算资源支持。

（三）数据的时效性

在疾病监测、实时医疗决策等医疗场景中，对数据的时效性要求较高。数据清洗需要在保证数据质量的前提下，尽可能快速地完成，以满足及时分析和应用的需求。

（四）数据隐私和安全

健康数据包含患者的个人敏感信息，如姓名、身份证号、疾病史等，在数据清洗过程中，需要严格遵守数据隐私与安全相关法律法规，采取数据加密、访问控制、匿名化等有效技术手段，充分保障患者数据的隐私和安全。

三、数据质量的提升方法

（1）数据验证。通过数据验证规则，确保输入数据符合预期格式和范围。

（2）数据清洗。定期对数据进行清洗，去除重复、错误和不一致数据。

（3）数据监控。建立数据质量监控机制，实时发现并纠正数据问题。

（4）培训和教育。对数据录入和处理人员进行培训，增强他们的数据质量意识。

素养之窗

肯尼亚 DHIS2 系统中的 HIV 指标数据清洗问题

在肯尼亚，DHIS2 系统被用于国家健康报告工作，包括 HIV 指标数据的收集、整理与分析。然而，在 2011 年至 2018 年的数据收集过程中，因缺乏有效的数据清洗流程，数据质量问题频发。这些问题包括数据重复、缺失值、错误值以及数据录入不规范等。尽管系统内置了一些数据质量检测机制，如数据验证规则和异常值分析，但这些机制在实际应用中未被充分激活，导致大量数据质量问题未被及时发现和纠正。这不仅影响了数据的完整性和准确性，还对基于这些数据的后续分析和决策造成了误导。

从该案例中可看出健康数据清洗在医疗数据管理中的重要性。数据清洗不仅是确保数据质量的关键步骤，更是保障医疗决策科学性和有效性的基础。缺乏有效的数据清洗流程，会导致数据重复、错误和缺失等问题，进而影响医疗研究的准确性和临床决策的可靠性。这提醒医疗机构和数据管理者必须重视数据清洗工作，建立健全的数据清洗机制，确保数据的完整性和准确性，从而为医疗研究和临床实践提供高质量的数据支持。

资料来源：GESICHO M B,WERE M C,BABIC A.Data cleaning process for HIV−indicator data extracted from DHIS2 national reporting system: a case study of Kenya.BMC Med Inform Decis Mak，2020,20:293.[2025−07−20].https://link.springer.com/article/10.1186/s12911−020−01315−7. DOI:10.1186/s12911−020−01315−7.

实战演练

针对本模块项目一“实战演练”所建立的“学生心理健康数据库”，开展数据清洗工作。具体操作涵盖处理缺失值、异常值以及重复数据，之后将清洗完成的数据导出，并将其命名为“学生心理健康数据库（清洗后）”，以便后续开展进一步分析。

任务二 筛选健康数据

任务描述


微课
数据筛选

本任务要求筛选出年龄在 60 ～ 80 岁、收缩压高于 140 mmHg 或舒张压高于 90 mmHg，且数据来源为社区卫生服务中心档案或现场体检的老人。本任务将通过

SPSS 软件，以老年人慢性病数据为例，帮助读者掌握数据筛选的基本操作技能，并使其能够根据实际研究需求灵活设置筛选条件，为后续深入分析老年人慢性病的流行趋势、风险因素及健康管理策略提供高质量的数据支持。

任务分析

本次任务所使用的数据源自经清洗的“老年人慢性病健康数据库”，该数据库包含老年人的姓名、性别、年龄、身份证号、血压指标以及数据来源渠道等健康相关信息，为任务的开展奠定了基础。筛选条件需通过逻辑运算符来设定，以便精准定位符合特定健康特征与数据来源要求的老年人群。可通过 SPSS 的“选择个案”功能完成此次数据筛选任务。

知识链接

一、数据筛选的概念

数据筛选是指根据预先设定的一个或多个条件，从原始数据集中选取符合条件的记录（或称为个案、样本、行等），同时排除不符合条件的记录，从而形成一个新的、规模更小但更具相关性和价值的数据子集的过程。这一过程类似用筛子过滤沙石，仅保留尺寸大于筛孔的颗粒，因此称为数据筛选。

二、数据筛选的重要性

在数据分析与处理过程中，数据筛选是一项至关重要的基础操作。它能够帮助我们从庞大的数据集中提取出满足特定条件的有用数据，为后续的深入分析、决策支持以及知识发现等环节提供精准且有针对性的数据子集。对数据筛选概念的深入理解，有助于更好地开展各类数据分析工作，充分发挥数据价值。

三、数据筛选的内容

（一）筛选变量（列）

可依据变量的名称、类型、属性等特征，选择特定变量进行筛选。例如，在一份包含多种数据类型（如数值型、字符型、日期型等）和不同主题（如客户信息、销售数据、产品参数等）的大型数据集中，若当前只关注与客户相关的分析，可以筛选出“客户编号”“客户姓名”“客户年龄”“客户所在地区”等客户相关变量，暂时忽略与销售、产品等无关的变量，从而聚焦于所需分析的客户维度数据。

（二）筛选记录（行）

基于记录中各变量的具体取值是否满足设定条件进行筛选。常见的筛选条件包括等于、不等于、大于、小于、大于等于、小于等于、介于特定范围之间、包含特定字符或模式等。以学生成绩数据表为例，若想找出所有数学成绩在 90 分及以上的学生，可以将筛选条件设置为“数学成绩 >=90”，然后数据筛选操作就会逐一检查每条记录的“数学成绩”字段值，保留符合条件的记录，形成包含高分学生的数据子集。

四、数据筛选的原则

（一）明确筛选目的

进行数据筛选前，必须清晰定义筛选目标，即通过筛选解决什么问题、得到什么样的数据结果。筛选目的的明确性将直接决定筛选条件的设置是否合理、精准。只有清楚知道要筛选出什么样的数据，才能避免盲目操作，防止遗漏重要信息或者筛选出大量无关数据，导致后续分析工作逻辑混乱。

（二）筛选条件的合理性

筛选条件的设定应基于对数据特征、业务逻辑以及分析需求的深入理解，确保其具有合理性和可行性。所设定的条件既不能过于宽松，导致筛选出的数据范围过大、包含大量冗余信息，增加后续处理的负担；也不能过于严格，导致筛选出的数据过少，无法满足分析所需的足够样本量，甚至得到缺乏统计代表性的结果。例如，在分析某产品在不同年龄段的市场接受度时，若将年龄段筛选条件设置过窄（如仅限 20 ～ 22 岁），可能因样本量过小无法准确反映该产品在整个目标市场中的真实情况；而若设置得过于宽泛（如包含所有年龄段），则可能无法凸显产品在特定年龄段的差异化表现。

（三）保持筛选条件的一致性

对同一数据集进行多次筛选，或者在不同时间点对相似的数据集进行筛选时，应尽量保持筛选条件的一致性。这有助于保证不同筛选结果之间的可比性，便于对数据在不同阶段或不同场景下的变化情况进行准确的分析和比较。例如，在定期跟踪企业客户流失情况时，每次进行数据筛选时都应采用相同的客户流失判定条件（如连续未购买时长、购买频率下降幅度等），才能客观地观察到客户流失趋势的变化，为制定客户挽留策略提供可靠依据。

五、SPSS 中的数据筛选方法

（1）简单筛选。根据单一条件筛选数据。

（2）复合筛选。根据多个条件组合筛选数据。

（3）临时筛选。筛选后仅在当前会话中生效，不影响原始数据。

（4）永久筛选。筛选后直接修改原始数据集。

任务实施

一、数据文件导入

启动 SPSS 软件，执行“File”→“Open”命令，在弹出的对话框中找到并选中“老年人慢性病健康数据库（清洗后）.xlsx”文件，打开并导入 SPSS 数据编辑器中。

微课

数据筛选的 SPSS 软件实现

二、筛选条件设置

步骤 1 执行“Data”→“Select Cases”命令，系统弹出“Select Cases”对话框，如图 2−52 所示。

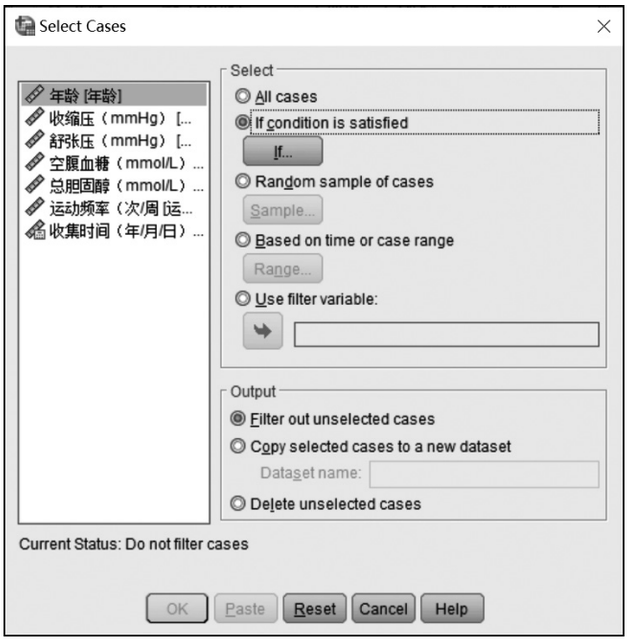


图 2-52 “Select Cases”对话框 4

步骤 2 设置年龄、血压、数据来源渠道的筛选条件。在“Select Cases”对话框中，选中“If condition is satisfied”单选按钮，单击“If”按钮，在弹出的“Select Cases: If”对话框中，输入“(年龄 >=60& 年龄 <=80)&(收缩压 >140| 舒张压 >90)&(数据来源渠道 =' 社区卫生服务中心档案 'OR 数据来源渠道 =' 现场体检 ')”，以筛选出“年龄在 60 ～ 80 岁”“收缩压高于 140 mmHg 或舒张压高于 90 mmHg ”且“数据来源渠道为社区卫生服务中心档案或现场体检”的记录，如图 2-53 所示。

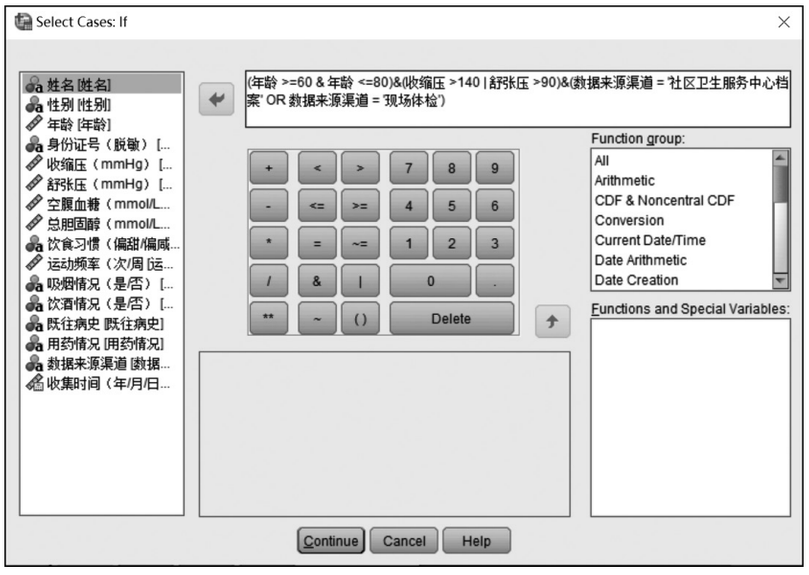
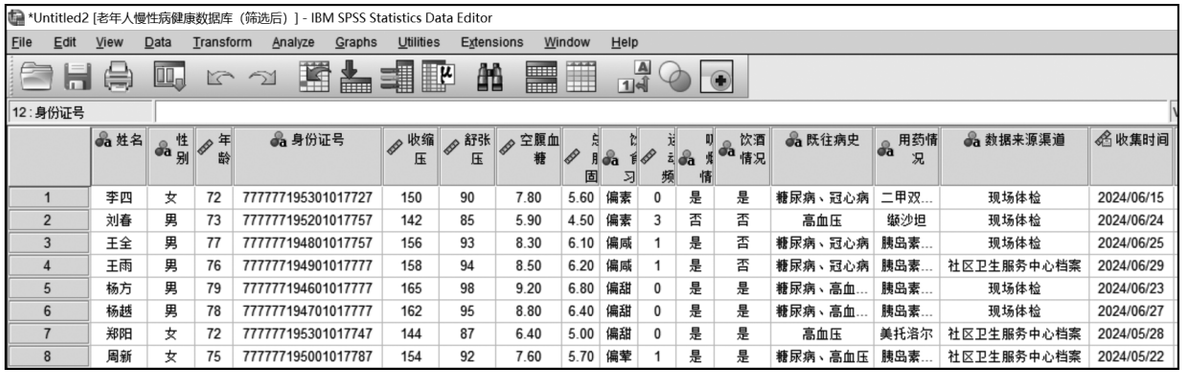


图 2-53 设置筛选条件

三、筛选结果验证

步骤 1 设置好所有筛选条件后，单击“Continue”按钮，选中“Copy selected cases to a new dataset”单选按钮，输入新数据集名称“老年人慢性病健康数据库（筛选后）”，再单击“OK”按钮，如图 2−54 所示。SPSS 将按设定条件进行数据筛选，并生成名称为“老年人慢性病健康数据库（筛选后）”的新数据集。筛选结果如图 2−55 所示。

图 2−54 将选定个案复制到新数据集



	姓名	性别	年龄	身份证号	收缩压	舒张压	空腹血糖	总胆固	饮酒习	运动频	吸烟情	饮酒情况	既往病史	用药情况	数据来源渠道	收集时间
1	李四	女	72	777777195301017727	150	90	7.80	5.60	偏素	0	是	是	糖尿病、冠心病	二甲双...	现场体检	2024/06/15
2	刘春	男	73	777777195201017757	142	85	5.90	4.50	偏素	3	否	否	高血压	缬沙坦	现场体检	2024/06/24
3	王全	男	77	777777194801017757	156	93	8.30	6.10	偏咸	1	是	否	糖尿病、冠心病	胰岛素...	现场体检	2024/06/25
4	王雨	男	76	777777194901017777	158	94	8.50	6.20	偏咸	1	是	否	糖尿病、冠心病	胰岛素...	社区卫生服务中心档案	2024/06/29
5	杨方	男	79	777777194601017777	165	98	9.20	6.80	偏甜	0	是	是	糖尿病、高血...	胰岛素...	现场体检	2024/06/23
6	杨越	男	78	777777194701017777	162	95	8.80	6.40	偏甜	0	是	是	糖尿病、高血...	胰岛素...	现场体检	2024/06/27
7	郑阳	女	72	777777195301017747	144	87	6.40	5.00	偏甜	0	是	是	高血压	美托洛尔	社区卫生服务中心档案	2024/05/28
8	周新	女	75	777777195001017787	154	92	7.60	5.70	偏荤	1	是	是	糖尿病、高血压	胰岛素...	社区卫生服务中心档案	2024/05/22

图 2−55 筛选结果

步骤 2 仔细检查筛选后的数据，查看其是否符合以下设定条件：年龄在 60 ～ 80 岁，且收缩压高于 140 mmHg 或舒张压高于 90 mmHg；数据来源渠道为社区卫生服务中心档案或现场体检；适用对象为老年人。如发现问题，可返回上一步骤，检查并修改筛选条件。

四、结果文件导出

确认筛选结果无误后，将筛选后的数据库保存并导出，以便后续进行深入的健康状况分析和管理。

知识拓展

一、数据筛选在实际应用中的作用

（一）提高分析效率

从海量数据中快速提取出与分析任务密切相关的数据子集，可降低数据处理的规模和复杂度，使分析师能够将更多时间和精力集中在关键数据的深入分析上，从而更高效地发现数据中的潜在规律和价值，为决策提供及时、准确的依据。

（二）保障数据质量

通过筛选剔除异常值、缺失值、重复值等不符合要求或存在问题的数据记录，能够有效提高数据的准确性、完整性和一致性，确保分析结果的可靠性，避免因数据质量问题得出错误的决策或误导性的结论。

（三）支持个性化分析

根据不同用户的需求和分析目标，灵活地设置筛选条件，提取特定的数据子集，可以满足多样化、个性化的数据分析需求，为不同部门、不同业务场景下的决策支持提供精准的数据服务。

（四）促进数据挖掘

为数据挖掘算法提供更优质的输入数据，在数据挖掘过程中，针对性强、质量高的数据子集，有助于提高挖掘模型的性能和准确性，更易发现隐藏在数据中的模式、关联规则和趋势等，进一步推动知识发现和创新。

二、数据筛选与数据分析方法的结合

（一）与统计分析方法的结合

数据筛选是统计分析的前置步骤，能够为统计分析提供精准的数据基础。例如，在进行数据分析中的相关性分析时，通过筛选符合特定条件的相关变量数据，可以更准确地计算变量之间的相关系数，揭示变量之间的内在关系。

以研究某地区的经济发展水平与居民生活质量之间的关系为例，可以先筛选出该地区不同时间段的经济指标数据（如 GDP、人均收入等）和居民生活质量指标数据（如预期寿命、教育水平、住房条件等），再运用统计分析软件（如 SPSS）进行相关性分析，从而深入了解经济发展对居民生活质量的影响程度，为政府制定区域发展政策提供科学依据。

（二）与数据挖掘技术的结合

数据挖掘技术旨在从大量数据中发现潜在的模式、知识和规律，而数据筛选在其中扮演着关键角色。在数据挖掘项目中，通常需要先对原始数据进行筛选，剔除无关数据与噪声数据，保留与挖掘目标相关的数据子集，以提高数据挖掘算法的效率和准确性。例如，在进行客户细分的数据挖掘任务时，可以先依据业务需求筛选出与客户特征和行为相关的数据，如消费金额、购买频率、产品类别偏好等，再运用聚类算法等数据挖掘技术对筛选后的数据进行分析，将客户划分为不同的群体，为企业精准营销、客户关系管理等提供有力支持。

素养之窗

医疗卫生领域专科专病高质量数据集

当前，医疗卫生行业数据正深陷“需求爆发与供给滞后”的结构性矛盾，面临数据孤岛、治理低效、隐私安全三重壁垒。中电四川数据服务有限公司基于城市级健康医疗数据基础设施，主动采集汇聚成都市近 700 亿条、43 T 临床诊疗数据，从海量的城市级临床诊疗数据中，结合大模型特性、医疗场景特征，精准筛选，形成了包含肺癌、肝癌、急性心肌梗死等 20 多个专病的专科专病高质量数据集及商保快赔主题高质量数据集，有效支撑了典型医疗场景落地。

资料来源：高质量数据集典型案例 | 医疗卫生领域专科专病高质量数据集 [EB/OL].(2025−11−20)[2025−11−26].https://www.nda.gov.cn/sjj/ywpd/szkjyjcss/1120/20251120134438807468715_pc.html.

实战演练

对本项目任务一“实战演练”中的“学生心理健康数据库（清洗后）”的数据进行筛选，选出年级为大三或大四且心理健康状态异常的学生信息。