

“正大杯”第十二届全国市场调查 与分析大赛公益培训

1.6 数据收集与整理

1.6 数据收集与整理

主要内容

1

目标群体的确定

2

问卷发放与数据收集

3

数据预处理

4

信度与效度分析



目标群体的确定

第十二届市湖大赛公益培训课件
(Credamo 见数版权所有)

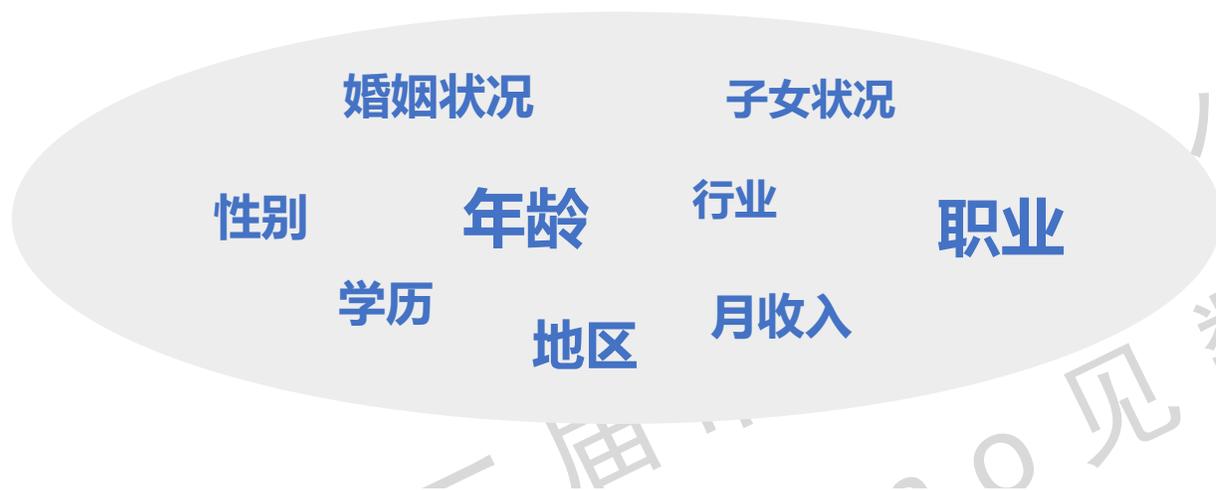
目标群体的确定

- 在收集数据前，首先要确定调研的目标群体
- 根据调研的目标与问题，对调研样本群体的特征进行设定
- 例如，对手机用户群体进行调研，可能需要设定目标群体的年龄（年轻人 or 老年人）、收入（高收入 or 低收入）、地域（一线城市 or 县域地区）等方面的特征
- 确定目标群体特征后，进行样本抽样



目标群体的确定

设定目标群体特征



样本特征设置  (1)最多选择两项, 限 [高级用户](#) (2)更多特征要求, 请 [联系我们](#)

人口统计

性别

年龄段

当前所在城市

行业

家庭情况

收入

婚姻状况

子女状况

[更多](#) ▼

- 在设定完目标被试特征后，Credamo平台会在样本数据库中，根据用户的数量需求，进行随机推送
- 目前Credamo平台的样本数据库中的被试超过200万，覆盖中国各个地区，以及部分海外地区

2

问卷发放与数据收集

第十二届市桥大赛公益培训课件
(Credamo 版权所有)

问卷发放与数据收集

- **传统市场调研**，以线下调研或纸质调研的方式展开，但存在多方面的缺点与不足：
 - 1) 受限于纸面调查的形式，调研所涉及的问题形式比较单一；
 - 2) 难以实现多期追踪调研、随机实验、显示逻辑等调研形式；
 - 3) 调研数据录入与清理困难，费时费力，并且容易出现错误；
 - 4) 传统市场调研的调研周期较长，成本较高。
- **网络市场调研**，又称在线市场调研，是指以互联网为载体和媒介，对消费者、公众、市场环境、竞争对手等进行信息数据进行调查分析的过程。
- 网络调研在题目类型、调研流程、质量控制、数据清理与分析等方面有着明显的优势。

问卷设计

问卷发布

数据清理

建模分析

新建发布

问卷发布名称 请输入问卷发布名称，每次发布可使用不同名称。

手机广告效果研究

*发布渠道 每次发布只能选择一种渠道

官方被试库

数据集市

200万+官方样本
千份问卷仅需1小时

- 基于Credamo平台被试的调研
- 可选人群标签，精准投放
- 快速回收，最低0.5元/份

需自行转发

匿名作答

作答者无需注册
不支持设置报酬

- 适用于对自己朋友圈的调研
- 自行转发问卷链接和二维码
- 作答者无需注册Credamo账户

需自行转发

链接/二维码

作答者需要注册
可自主设置报酬

- 适用于自有样本的有偿调研
- 自行转发问卷链接和二维码
- 可设置报酬和随机红包

追踪调研

样本库

对已作答样本进行追踪调研

- 适用于纵向或跨期研究
- 可在数据清理中导入被试
- 对选定被试定向发放问卷

企业专享

企业联系人

导入联系人后定向调研

- 适用于企业对内部人员进行调研
- 仅限企业版用户

问卷发放与数据收集

• 质量控制手段

- 作答经验
- 信用分
- 采纳率
- IP限制
- 过滤指定问卷被试
- 限制作答设备
- 作答区域范围控制
- 限制重复作答

质量控制 ? 可有效提高被试质量和 问卷作答质量 限入门版及以上用户

被试已作答次数

被试信用分

被试历史采纳率

过滤指定问卷用户

*选择指定问卷，本次发布将禁止作答过指定问卷的用户作答

仅限电脑端作答。 (如问卷需播放音视频，请务必勾选)

*勾选后，被试只能通过电脑端作答。

作答区域范围

*勾选后在该范围区域内仅允许1人作答。

智能行为式验证

*作答前进行智能人机验证，大幅提高数据质量和安全，任选一种。

问卷发放与数据收集

- 样本匹配

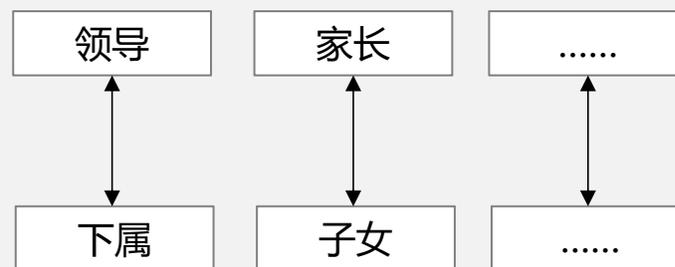
样本匹配

应用描述：
流程化实现“1对1”配对样
本调查

免费

 credamo

打开



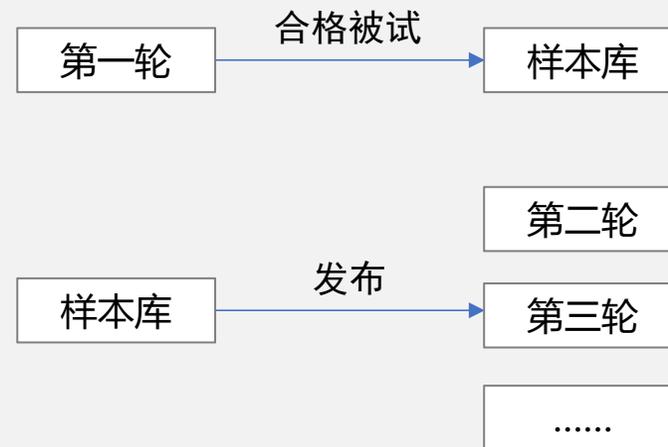
问卷发放与数据收集

- 多期追踪



○ 样本库

可在数据清理中，导入被试列表；
针对特定的被试列表发放问卷；
实现对相同被试的多期跟踪；
适合纵向或跨期研究。



3

数据预处理

第十二届市调大赛公益培训课件
(Credamo 版权所有)

数据预处理

- 对收集的数据进行初步的整理，以达到分析的要求

The interface shows a workflow with four steps: 问卷设计 (Questionnaire Design), 问卷发布 (Questionnaire Release), 数据清理 (Data Cleaning), and 建模分析 (Modeling Analysis). The '数据清理' step is currently active. Below the workflow, there are three tabs: 数据概览 (Data Overview), 变量设置 (Variable Settings), and 下载管理 (Download Management). The '变量设置' tab is selected, showing a search bar for variable names and four action buttons: 初始化变量 (Initialize Variable), + 添加变量 (Add Variable), 按文本显示 (Display by Text), and 按序号显示 (Display by ID). Below these is a table with 6 rows and 4 columns: 序号 (Serial Number), 标题 (Title), 变量名 (Variable Name), and 显示方式 (Display Method). Each row contains a variable name and its corresponding display method, which is currently set to '选项文本显示' (Option Text Display).

序号	标题	变量名	显示方式
1	作答ID	作答ID	选项文本显示
2	用户ID	用户ID	选项文本显示
3	用户名	用户名	选项文本显示
4	开始时间	开始时间	选项文本显示
5	结束时间	结束时间	选项文本显示
6	作答时长(秒)	作答时长	选项文本显示

数据预处理

- 初始化变量



在收集完数据或是修改完问卷后，需要对变量进行初始化操作，否则可能出现数据显示不全的情况。

注意：初始化变量后，原有自定义的变量会被删除，需要重新添加；原有的建模分析结果也需要重新进行分析

数据预处理

- 添加变量

The screenshot shows the '数据清理' (Data Cleaning) step in the Credamo software. A red arrow points to the '+ 添加变量' (Add Variable) button. Below it, a dialog box titled '添加变量' (Add Variable) is open. The dialog has a search bar for '新变量名称' (New Variable Name) with '信息性' (Informational) entered. It features a '普通表达式' (Normal Expression) section with the formula $(var1 + var2 + var3)/3$ and a '变量列表' (Variable List) section containing three entries: 'var1 = 信息性1', 'var2 = 信息性2', and 'var3 = 信息性3'. At the bottom, there are '添加变量' (Add Variable), '删除变量' (Delete Variable), '取消' (Cancel), and '确定' (Confirm) buttons.

在原有数据的基础上，通过计算得到新的变量。例如，案例问卷中“信息性”、“趣味性”等心理学变量。

或者问卷中获得了身高、体重，需要得到BMI指数

数据预处理

- 按文本显示

The screenshot shows a software interface for data management. At the top, there are four main steps: '问卷设计' (Questionnaire Design), '问卷发布' (Questionnaire Release), '数据清理' (Data Cleaning), and '建模分析' (Modeling Analysis). Below these, there are three sub-steps: '数据概览' (Data Overview), '变量设置' (Variable Settings), and '下载管理' (Download Management). A search bar is present with the placeholder text '请输入变量名称' (Please enter variable name). To the right of the search bar are four buttons: '初始化变量' (Initialize Variable), '+ 添加变量' (Add Variable), '按文本显示' (Display by Text), and '按序号显示' (Display by Number). A red arrow points to the '按文本显示' button. Below the interface is a table of survey data.

-我认为这个广...	-我认为这个广...	-我认为这个广...	我认为这个广告...	请您根据刚看到...
同意	有点同意	同意	1	1
有点不同意	不同意	不同意	3	1
不同意	不同意	不同意	3	1
有点同意	不同意	有点不同意	3	1
有点同意	同意	非常同意	3	1
同意	有点同意	非常同意	3	1
中性	不同意	中性	3	1
不同意	不同意	不同意	3	1
有点同意	有点同意	同意	3	1
有点同意	有点同意	同意	3	1

✓ 数据将以选项内容的形式呈现

✓ 此时下载的数据也将以选项文本呈现

数据预处理

- 按序号显示



✓ 数据将以选项序号的形式呈现

✓ 此时下载的数据也将以选项序号呈现

-我认为这个广...	-我认为这个广...	-我认为这个广...	我认为这个广告...	请您根据刚看到...
6	5	6	1	1
3	2	2	3	1
2	2	2	3	1
5	2	3	3	1
5	6	7	3	1
6	5	7	3	1
4	2	4	3	1
2	2	2	3	1
5	5	6	3	1
5	5	6	3	1

4

信度与效度分析

第十二届市桥杯大赛公益培训课件
(Credamo 见数版权所有)

信度与效度分析

• 信度分析

信度 (Reliability) 即可靠性, 它是指采用同样的方法对同一对象重复测量时所得结果的一致性程度, 通常用于对心理学变量的不同问题的测度进行分析 (如李克特量表等)

【维度1】Cronbach' α 信度分析表

下载数据

变量名称	CITC (校正的题项与总体相关性)	题项删除后的Cronbach' α 系数	Cronbach' α 系数
信息性1	0.6259	0.7557	
信息性2	0.7014	0.6674	0.791
信息性3	0.6448	0.7031	

智能解读

- 【维度1】整体的Cronbach' α 系数为0.791, 说明各题项的内部一致性较好, 数据信度比较理想。
- 综上所述, 【维度1】的总体数据信度达标, 可用于后续分析。

信度与效度分析

• 效度分析

效度 (Validity) 即准确性, 指理论变量与测量维度之间的对应关系, 可以通过探索性因子分析或验证性因子分析进行分析

添加分析 ×

分析方法 分析名称

效度分析仅支持数值变量

维度1

作答时长 信息性1

信息性2 信息性3

信度与效度分析

- 效度分析（续）

KMO检验与Bartlett球形检验

下载数据

指标	数值
样本量	80
KMO值	0.9064
Bartlett球形值	619.4939
Bartlett球形检验自由度	79
Bartlett球形检验P值	0

信度与效度分析

• 效度分析 (续)

因子载荷系数与共同度

下载数据

维度	题项	因子载荷系数				共同度
		因子1	因子2	因子3	因子4	
维度1	信息性1	0.1823	0.7785	0.164	0.232	0.72
	信息性2	0.2759	0.8246	0.1737	-0.068	0.7908
	信息性3	0.2175	0.7174	0.293	0.0958	0.6571
维度2	娱乐性1	0.825	0.0539	0.3644	-0.066	0.8207
	娱乐性3	0.7833	0.4193	0.0703	0.195	0.8324
	娱乐性2	0.7191	0.2293	0.2157	0.3264	0.7227
维度3	注意力1	0.8584	0.2728	0.1126	0.226	0.8751
	注意力2	0.8177	0.3101	0.1699	0.1865	0.8285
	注意力3	0.2645	0.0992	0.1293	0.9116	0.9276
维度4	说服力1	0.5138	0.3	0.6004	0.004	0.7145
	说服力3	0.3754	0.5362	0.4643	0.0478	0.6463
	说服力2	0.1541	0.3517	0.8055	0.2208	0.845

信度与效度分析

- 效度分析（续）

旋转后特征根值与方差贡献率

下载数据

指标	因子1	因子2	因子3	因子4
旋转后特征根值	6.3684	1.0855	0.7355	0.4551
方差贡献率 (%)	32.27	22.61	13.42	9.87
累计方差贡献率 (%)	32.27	54.88	68.3	78.17

信度与效度分析

• 效度分析（续）

智能解读

- 问卷的效度指的是问卷能在多大程度上反映它所测量的理论维度，而结构效度是指测量题项与测量维度之间的对应关系符合预期的程度，一般通过0.9064，Bartlett球形检验 $P = 0 < 0.05$ ，说明数据很适合因子分析。
- 另外，旋转后累计方差贡献率为7.8172%，说明因子分析有效地提取了数据的信息。
- 样本量过低（ $n = 80$ ）。一般的量表问卷建议样本量至少为题项数 * 10（即120），另外进行效度分析建议样本量至少要为100。
- 维度1、维度2、维度3、维度4包含题项过少。一般建议一个维度设置4-7个变量。若您正在进行预测试，建议修改问卷，增加相应维度的题项。
- 若题项与某因子的因子载荷系数绝对值大于0.4，可以认为该题项与该因子之间存在对应关系。理想情况下，一个题项仅与一个因子存在对应关系，度最为理想。您可以通过删除对应关系不合理的题项来提升问卷效度。
- 从上表可知，维度与因子间对应关系不符合预期，问卷效度较差。
- 维度2、维度3与因子1、因子4对应关系不明确，建议先处理其他维度，或确定题项与维度是否正确对应、增大样本量，以及重新设计量表。

谢谢大家

关注右边公众号



及时了解大赛资讯和进程

随时学习大赛公益培训

中国商业统计学会

官方网址：<http://www.china-cssc.org>

公众号：



Credamo见数

官方网址：www.credamo.com

公众号：

