第二题、2015 年北师大赛 A 题：

问卷调查的可靠性分析实例

——中国人主观幸福感综合评价

1 问题的重述与分析

问题要求给出依据数据判定各因素对主观幸福感影响的重要性的模型和具体的算法，并针对不同群体分析主要因素。

通过分析数据说明，我们认为数据中的不同变量存在逻辑上的因果关系，因此可以将“主观幸福感”视为一个多元函数 g 的因变量，各项满意程度（例如“满意程度-家庭经济情况”等）为中间变量，而影响各满意程度的因素为自变量，从而问题变成了构建主观幸福感与各自变量之间关系的数学模型，并将该模型应用到不同群体中的问题。

我们定义各因素的重要性为自变量影响因变量的程度，即该自变量在多元函数 g 中的指数与系数的大小。指数越大的自变量对因变量影响程度越大，即越重要，例如在 g 函数表达式中为 3 次方的自变量重要性大于 2 次方的自变量。当指数相同时，则通过比较自变量前系数判断重要性大小，系数越大越重要。

首先，通过数据同向化、结构化等手段对数据预处理，再进行可靠性分析，运用箱形图方法和拉依达准则法对数据去噪，从而得到有价值的子数据集。

其次，在去躁后的数据集中，对所有样本不进行分类，依次对各自变量与相关的中间变量、各中间变量与主观幸福感进行多元回归分析，得到因变量的函数

表达式。在函数表达式中指数和系数越大的自变量，越重要。以此为依据将所有自变量按照重要性排序，此时，我们得到了对于所有群体影响“主观幸福感”各因素重要性的序列。

之后，对不同的定类量表（例如性别、党派、年龄段）进行控制变量，将所有样本分为不同群体。针对其中的每一个群体，再次重复上述的分析过程，生成影响该群体“主观幸福感”的各因素重要性序列。其中，在序列前端的因素即为影响该群体“主观幸福感”的主要因素。

关键词： 可靠性分析 数据去噪 多元回归分析

2 假设与理由

1) 假设该调查覆盖面足够广，样本空间多样性足够大，抽样随机性足够好。否则无法排除由于样本元素集中于某一特定区域（例如大城市、汉族聚集地）而引起的误差。

2) 假设定类量表（性别、党派等）准确。结合实际，问卷调查中个人基本信息与现实不符的可能性极低，因此忽略其误差。

3) 假设各自变量平等，且对因变量的影响程度相同，即没有权重。在数据预处理中，由于无法确定哪些自变量对因变量有更大影响力，因此可以先做无权重假设，之后通过具体的数据分析，再为每一个自变量设置权重。

4) 假设对某一量表作出“不适用\不好说\不知道\拒绝回答\不作选择”回答的被调查者，其真实情况服从均匀分布，例如，在“个人总收入”中回

答“拒绝回答”的被调查者，其实际情况既可能是总收入高，也可能是总收入低，且两者等可能出现。

5) 假设该调查无定值系统误差，忽略调查时的环境、问卷的设计等原因造成的整体结果的单一偏向性，即调查结果中没有因变量整体偏低或偏高的情况（例如，测试幸福感相对于真实幸福感整体偏低）。但存在随机误差或者非定值系统误差，即调查员对数据的误操作，或调查有时在让人愉悦的环境中进行（例如，高档娱乐场所）而有时在让人厌烦的环境中进行（例如，闷热的公交站）。

3 数据来源

中国综合社会调查（Chinese General Social Survey，缩写为 CGSS）

4 数据初步描述

该问卷全部是单选题的形式。

调查数据以定类量表与定序量表为主，有少数定距量表。

各定类量表（如性别、党派、年龄段等）值域完整，分布均匀。

各定序量表（如幸福感、满意程度、收入认可等）间因果关系明确，结构清晰，且聚合性较好。

5 对数据的理解

定类量表不存在可靠性问题，我们假设调查中每个人的性别、党派、户籍等是准确的（假设 2）。但可以由此划分不同群体，构成样本空间的子空间，之后可以在群体的子空间中分析其他定序量表的可靠性。

定序量表中部分与态度、观点、满意程度等有关的变量，则是可靠性分析的主要对象。理论上，该调查中所有定序量表都与幸福感有相关关系，但相关性存在较大差异。同时，可以注意到其中有部分变量可以作为中间变量，例如与满意程度有关的变量。例如变量“家庭经济状况满意程度”，既可以视为因变量“主观幸福感”的自变量，但又可以视为自变量“家庭的社会经济地位”的因变量。根据这种因果关系可以将量表结构化，便于分层次的数据分析。

而定距量表则可以转化为定序量表来处理。

6 数据的标准化处理（预处理）

6.1 缺失数据处理

在分析相关数据时，删去带有缺失数据的抽样。

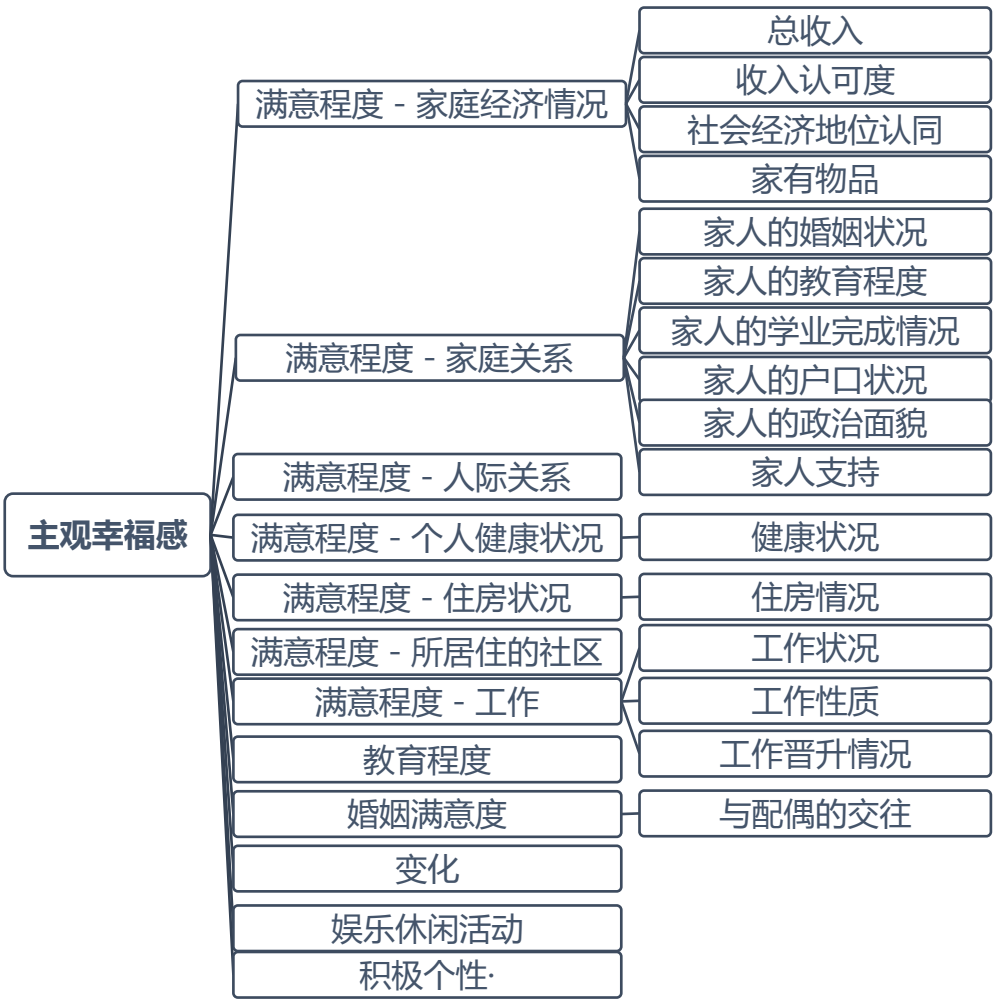
6.2 无效数据处理

注意到调查问卷中有“不适用\不好说\不知道\拒绝回答\不作选择”的选项，考虑到每个人出于个人隐私的忧虑或者其他各方面原因而不愿透露，同时根据之前的均匀分布假设（3），可以认为无从得知这些数据与因变量之间的相关关系，因此定义其为无效数据，处理方法同缺失数据。

6.3 同向化处理

从数据说明中我们看到，在幸福感与各项满意度中，“非常幸福\非常满意”被设置为 1，而相应的，“很不幸福\很不满意”被设置为 4。但是有部分数据是反向编码，例如，对婚姻的满意程度中，“非常满意”被设置为 4，“很不满意”对应 1。为了数据分析的便利，我们将所有数据同向化，即从 1-4 分别对应 “非常满意\比较满意\不太满意\很不满意”。

6.4 结构化处理



结合实际推断各变量之间是否有因果关系，并根据因果关系建立树状结构图。图中的叶子对应一个内部关联性强的子变量集，叶子的根节点对应该子变量集中所有自变量的共同因变量。例如，图中根节点“满意程度-家庭经济情况”是他的叶子中各自变量“总收入\收入认可度\社会经济地位认同\家有物品”的因变量。之后我们将以子变量集为单位进行数据的可靠性分析。

6.5 定序化处理

该调查中大部分数据都可以归类为定序量表，但仍存在部分定距量表（例如总收入，家中书籍数量）。在分析各自变量的内部一致性时，出于数据标准化考虑，我们将所有数据根据区间分类定序化，例如，将收入分为 1-5 类，分别“对应高收入\较高收入\中等收入\较低收入\低收入”。

6.6 李克量表化处理

对于能够对同一因变量产生影响的多个定序自变量，通过加和构造一个新的李克量表，来分析各个自变量的可靠性以及内部一致性。例如，对于对幸福感有直接相关的各项满意程度，根据自变量平等假设（2），可以将这些满意程度变量无权重加和，构造一个新的中间变量：“满意程度总分”。

7　可靠性分析

在数据的预处理中，我们已经将所有变量通过结构化处理分成内部关联性强的子变量集，下面我们将以其中一个子变量集“家人支持”为例进行可靠性分析。由于篇幅限制，其余子变量集可通过同样方法分析。

家人支持	qhc2a	hc2a. 过去一年，您是否经常为自己父母提供帮助 - 给钱
	qhc2b	hc2b. 过去一年，您是否经常为自己父母提供帮助 - 帮助料理家务或照顾小孩或其他家人
	qhc2c	hc2c. 过去一年，您是否经常为自己父母提供帮助 - 听他(们)的心事或想法
	qhc3a	hc3a. 过去一年，您自己父母是否经常为您提供帮助 - 给钱
	qhc3b	hc3b. 过去一年，您自己父母是否经常为您提供帮助 - 料理家务
	qhc3c	hc3c. 过去一年，您自己父母是否经常为您提供帮助 - 听他(们)的心事或想法
	qhc5a	hc5a. 过去一年，您对跟您最亲近的成年子女是否经常提供帮助 - 给钱
	qhc5b	hc5b. 过去一年，您对跟您最亲近的成年子女是否经常提供帮助 - 料理家务
	qhc5c	hc5c. 过去一年，您对跟您最亲近的成年子女是否经常提供帮助 - 听他(们)的心事或想法
	qhc6a	hc6a. 过去一年，跟您最亲近的成年子女是否经常为您提供帮助 - 给钱
	qhc6b	hc6b. 过去一年，跟您最亲近的成年子女是否经常为您提供以下帮助 - 料理家务
	qhc6c	hc6c. 过去一年，跟您最亲近的成年子女是否经常为您提供帮助 - 听他(们)的心事或想法
	qhc7a	hc7a. 过去一年，您是否经常为您配偶父母提供帮助 - 给钱
	qhc7b	hc7b. 过去一年，您是否经常为您配偶父母提供帮助 - 帮您料理家务
	qhc8a	hc8a. 过去一年，您配偶父母是否经常为您提供帮助 - 给钱
	qhc8b	hc8b. 过去一年，您配偶父母是否经常为您提供帮助 - 帮您料理家务

对于子变量集“家人支持”，根据预处理中李克量表化的结果，我们已有一个新的定序量表“家人支持总分”，该量表是由“家人支持”中所有变量值加和得到。经过标准化处理，删去“家人支持”子变量集中的缺失数据与无效数据，仍保留 381 个数据项。

7.1 可靠性定义

在可靠性分析之前，我们对可靠性进行如下定义：

问卷选项可靠性：子变量集中各自变量对应的问卷选项之间具有显著的同质性。通过决断值和内部一致性α系数检验。由于子变量集中，各自变量是对共同因变量进行测验，因此当这些选项之间具有较好的内部一致性时，能够可靠的反应不同自变量对因变量的影响。相反，对不具有显著同质性的变量，或者不属于

该子变量集，或者与因变量相关关系较弱。出于简化问卷的考虑，应该删除该变量对应的选项。

逻辑可靠性：个体问卷数据中各项答案具有逻辑一致性。通过预设的逻辑判断算法检验。例如，各项满意程度中均选择“非常满意”的问卷，在主观幸福感中选择“很不幸福”，则可以视为不具有逻辑一致性。

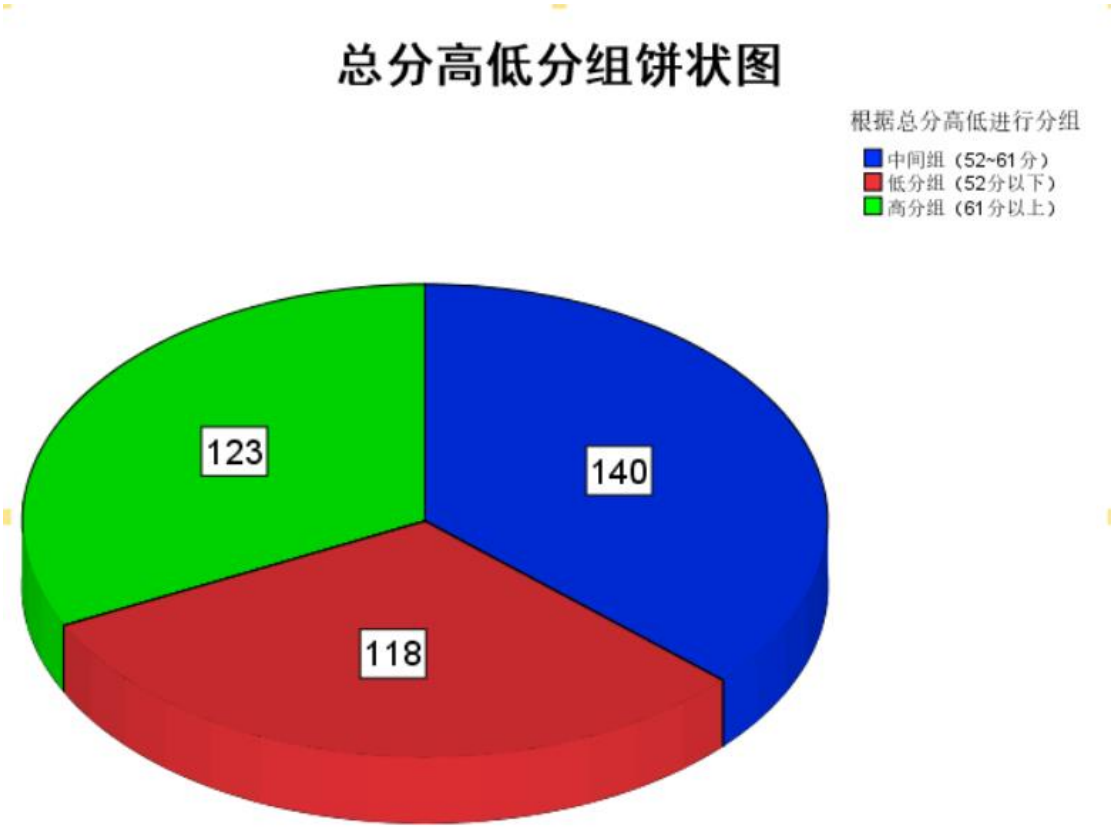
数据可靠性：数据集中不存在严重偏离整体的离群值。通过箱形图或拉依达准则法检验。例如，个人总收入中为“9999999999999”的数据应视为离群值。

7.2 问卷选项可靠性分析

7.2.1 决断值分析

决断值，即根据李克量表化的总分区分出高分答卷者与低分答卷者后，用于衡量高、低两组答卷者每个问卷选项得分的平均数差异的显著性的指标，其原理等同于独立样本的 t 检验。在本例中，高分答卷者即“家人支持总分”数值处于样本空间高 27%的答卷者，而低分答卷者则为“家人支持总分”数值出于低 27%的答卷者。决断值表达了问卷选项的鉴别度，其主要目的在于判别试题是否具有区别答卷者的目的。例如，当一个选项的对应的自变量的值，无论在高分答卷者，还是在低分答卷者中都具有相同的分布，那么该选项并不能很好的区分答卷者，可以考虑从选项中删除，或者分析其原因，将其纳入其他子变量集。

下图为对“家人支持总分”进行高、低分组的饼状图：



对于其中的每一个选项，首先通过 Levence 检验法判断高、低分组的方差是否相同，称为方差齐性检验。当假设检验概率 p1<0.05 时，应拒绝虚无假设：H_0：D1=D2，接受对立假设：H_0：$D1 \neq D2$，即两者方差不相等。而根据方差相等与否，对应的独立样本 t 检验计算方式也有所不同。独立样本 t 检验用来判断两个样本平均数与其各自所代表的总体的差异是否显著。这里我们用 t 检验的显著性概率 p2 衡量决断值，当 p2<0.5 时，达到 0.5 显著水平，即高、低分组中同一变量存在显著平均数差异，该问卷选项有效的区分了答卷者，反之则不能区分答卷者。

下图为 t 独立样本检验的计算结果表格：

独立样本检验

		Levene 的方差齐性检验		平均数相等的t检验	
		F	显著性概率	T	显著性概率
qhc2a	假设方差相等	4.867	.028	-9.053	.000
	不假设方差相等			-9.077	.000
qhc2b	假设方差相等	5.038	.026	-11.549	.000
	不假设方差相等			-11.605	.000
qhc2c	假设方差相等	.859	.355	-11.827	.000
	不假设方差相等			-11.851	.000
qhc3a	假设方差相等	77.083	.000	-9.655	.000
	不假设方差相等			-9.531	.000
qhc3b	假设方差相等	20.992	.000	-10.891	.000
	不假设方差相等			-10.802	.000
qhc3c	假设方差相等	.012	.914	-11.410	.000
	不假设方差相等			-11.384	.000
qhc5a	假设方差相等	9.371	.002	-6.462	.000
	不假设方差相等			-6.482	.000
qhc5b	假设方差相等	12.844	.000	-9.353	.000
	不假设方差相等			-9.404	.000
qhc5c	假设方差相等	10.933	.001	-9.056	.000
	不假设方差相等			-9.096	.000
qhc6a	假设方差相等	3.491	.063	-5.008	.000
	不假设方差相等			-4.992	.000
qhc6b	假设方差相等	.925	.337	-11.815	.000
	不假设方差相等			-11.787	.000
qhc6c	假设方差相等	.124	.725	-11.026	.000
	不假设方差相等			-11.025	.000
qhc7a	假设方差相等	.142	.707	-11.185	.000
	不假设方差相等			-11.180	.000
qhc7b	假设方差相等	2.383	.124	-13.406	.000
	不假设方差相等			-13.424	.000
qhc8a	假设方差相等	144.201	.000	-11.255	.000
	不假设方差相等			-11.075	.000
qhc8b	假设方差相等	42.081	.000	-13.292	.000
	不假设方差相等			-13.149	.000

可见，对于所有问卷选项，无论方差是否齐性，t 检验显著性概率 p2 均等于 0，且 T 绝对值大于一般标准值 3.00，即所有选项均有显著决断值，能够较好区分答卷者。

7.2.2 信度分析

信度，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，定义为真实分数的方差占测量分数的方差的比例。信度越大，则测试结果越可靠。在本例中，我们采用常见的克隆巴赫α系数对子变量集进行信度估计，如果该问卷选项删除后的克隆巴赫α系数比未删除前高出许多，则该选项与其余选项所要测验的性质可能不同，可以考虑删去，或进一步分析其原因。

首先，计算所有问卷选项保留时的克隆巴赫α系数，根据公式：

$$\alpha = \frac{p^2\bar{\sigma}_{ij}}{\sigma_x^2} = \frac{p}{p-1}\left(1 - \frac{\sum_{j=1}^{p}\sigma_j^2}{\sigma_x^2}\right)$$

其中 p 为选项总数，σ_j^2 表示第 j 个选项的方差，σ_x^2 表示李克量表“家人支持总分”的方差，通过计算可得克隆巴赫α系数为 0.831。

下一步，对每个选项，计算该选项对应变量与“家人支持总分”的积差相关，即描述两者的相关系数。该系数越大，则表示该选项与其余选项同质性高，若该系数小于 4，则同质性不高，可以考虑删去。计算公式如下：

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}}$$

X,Y 分别表示该选项与“家人支持总分”，N 表示样本数。

最后，计算删去该选项后的克隆巴赫α系数，公式同上。

下图为各选项与总分的积差相关，以及删去后的克隆巴赫α系数计算结果表格：

信度分析

	该选项与总分的积差相关	Cronbach Alpha（如果该选项已删除）
qhc2a	.400	.824
qhc2b	.465	.820
qhc2c	.515	.818
qhc3a	.487	.820
qhc3b	.469	.820
qhc3c	.492	.819
qhc5a	.214	.837
qhc5b	.375	.827
qhc5c	.428	.823
qhc6a	.211	.836
qhc6b	.500	.818
qhc6c	.490	.819
qhc7a	.497	.819
qhc7b	.518	.817
qhc8a	.544	.817
qhc8b	.563	.815

可见，仅有选项 qhc5a 与 qhc6a 的相关系数小于 4，且删去后克隆巴赫α系数提高。查阅数据说明，这两选项对应内容如下：

qhc5a：过去一年，您对跟您最亲近的成年子女是否经常提供帮助 - 给钱

qhc6a：过去一年，跟您最亲近的成年子女是否经常为您提供帮助 - 给钱

巧合的是，这两个选项都与家人之间的金钱互助有关，而我们的信度分析同时也说明，给钱在“家人支持”中的同质性较弱，即给钱在家人支持的各种形式中处于一个较不重要的地位，不能很好地反应出家人支持的总体状况。

7.3 逻辑可靠性分析

在“家人支持”中，并没有明显逻辑矛盾的选项，因此我们从整个数据集的角度进行逻辑可靠性分析。通过分析数据说明，我们发现逻辑矛盾主要集中于工作、教育和婚姻三类选项中，下面将逐一进行逻辑可靠性分析。

7.3.1 逻辑可靠性分析：工作类选项

工作状况	qb01b	那您目前的状况是什么？	1=正在上学、参军/服兵役	2=从未工作过	3=失去工作后正在找工作	4=休长假	5=离/退休	6=在家	7=丧失劳动能力	8=下岗	9=年迈	10=其他
	qb01c	您目前从事什么类型的工作？	1=全日工作	2=非全日工作	3=临时性工作（打零工/散工；非稳定性工作）							
工作性质	qc09_2	是否经常遇到需要繁重的体力劳动的情况	1=总是	2=经常	3=又是	4=很少	5=从不					
	qc09_4	是否经常遇到需要快速反应的思考或脑力劳动的情况	1=总是	2=经常	3=又是	4=很少	5=从不					
工作晋升情况	qc28	在过去的三年内，您是否获得过技术等级或职务上的晋升？	1=是	2=否								
	qc29	在过去的三年内，您是否获得过工资等级上的晋升？	1=是	2=否								
	qc30	在未来的几年内，您在单位里得到提拔或升迁的机会有多大？	1=几乎肯定会	2=很有可能	3=不太可能	4=几乎不可	5=不适用	6=不知道/不好说				
	qc31	如果换单位的话，您的提拔、升迁机会有多大？	1=几乎肯定会	2=很有可能	3=不太可能	4=几乎不可	5=不适用	6=不知道/不好说				

显然，如果 qb01b 选项非空，则表示答卷者并未工作，其余选项均为无效值。可根据如下逻辑表达式筛选出无效数据，SYSMIS（x）表示该个体的 x 变量为空，~，&，|分别为逻辑运算符非，与，或。

~ SYSMIS(qb01b) & (~ SYSMIS(qb01c) | ~ SYSMIS(qc09_2) | ~ SYSMIS(qc09_4) | ~ SYSMIS(qc28) | ~ SYSMIS(qc29) | ~ SYSMIS(qc30) | ~ SYSMIS(qc31))

通过该方法可筛选出 2224 个无效数据。

7.3.2 逻辑可靠性分析：教育类选项

教育程度	qa05a	您目前的最高教育程度是	1=没有受过	2=扫盲班	3=小学	4=初中	5=职高	6=普高	7=中专	8=技校	9=大专（成人）	10=大专（正规）	11=大学本科（成人）	12=大学本科（正规）	13=研究生	14=其他
	qa05d	您一共受过多少年的学校教育呢？	97=不适用	98=不知道	99=拒绝回答											

显然，如果 qa05a 选项值为 1，则 qa05d 为无效值。可根据如下逻辑表达式筛选：

最高教育程度 = 1 & ~SYSMIS(受教育年限)

通过该方法可筛选出 0 个无效数据。

7.3.3 逻辑可靠性分析：婚姻类选项

与配偶的交往	qd01	d1. 您目前的婚姻状况属于以下哪-	1=从未结过婚	2=同居	3=已婚有配偶	4=分居	5=离婚	6=丧偶	
	qhh01a	hh1a. 配偶会听我说我的烦恼	1=非常符合	2=相当符合	3=有些符合	4=无所谓符合不符合	5=有些不符合	6=相当不符合	7=非常不符合
	qhh01b	hh1b. 配偶会跟我说他/她的烦恼	1=非常符合	2=相当符合	3=有些符合	4=无所谓符合不符合	5=有些不符合	6=相当不符合	7=非常不符合
	qhh06	hh6. 若是有机会再次选择您的配偶	1=一定会	2=大概会	3=大概不会	4=一定不会			

显然，如果 qd01 选项值为 1\5\6，则 qhh01a\qhh01b\qhh06 均为无效值。课根据如下逻辑表达式筛选：

(婚姻状况 = 1 | 婚姻状况 = 5 | 婚姻状况 = 6) & (~ SYSMIS(配偶是否倾听) | ~ SYSMIS(配偶是否倾诉) | ~ SYSMIS(是否重新选择配偶))

通过该方法可筛选出 0 个无效数据。

综上，该数据集中，除了工作类选项的数据逻辑矛盾率较大以外（23.36%），其他选项均有较好的逻辑一致性。

7.4 数据可靠性分析与数据去噪

7.4.1 箱形图方法

7.4.1.1 绘制箱形图

箱形图，提供了一种只用 3 个点对数据集做简单总结的方式，常用来筛选偏离中心较大的异常值。这 3 个点包括中位点、Q1（上四分位点）、Q3（下四分位点）。箱形图很形象的分为中心、延伸以及分布状态的全部范围。

箱形图的绘制步骤：

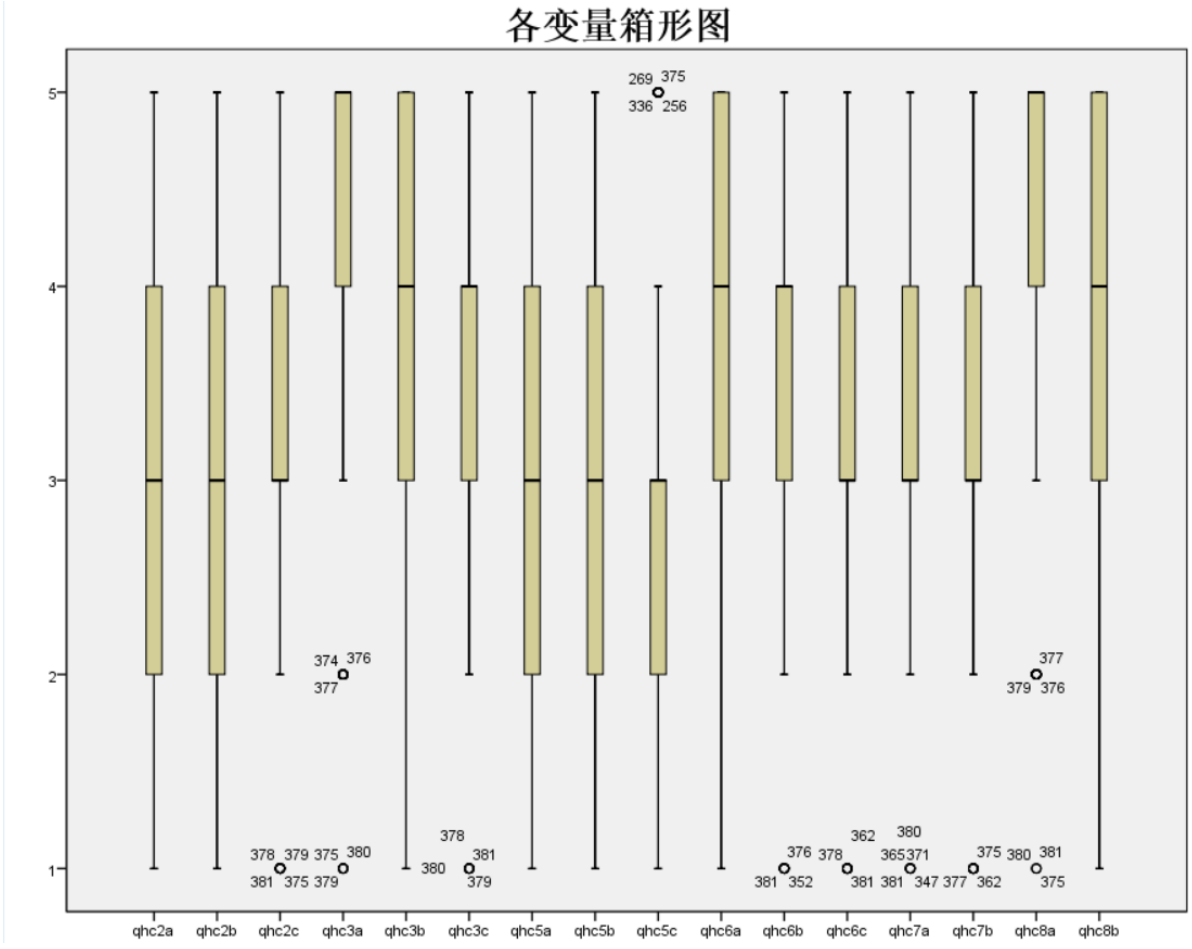
1、画数轴，度量单位大小和数据的单位一致，起点比最小值稍小，长度比该数据批的全距稍长。

2、画一个矩形盒，两端边的位置分别对应数据批的上下四分位数（Q1 和 Q3）。在矩形盒内部中位数（Xm）位置画一条线段为中位线。

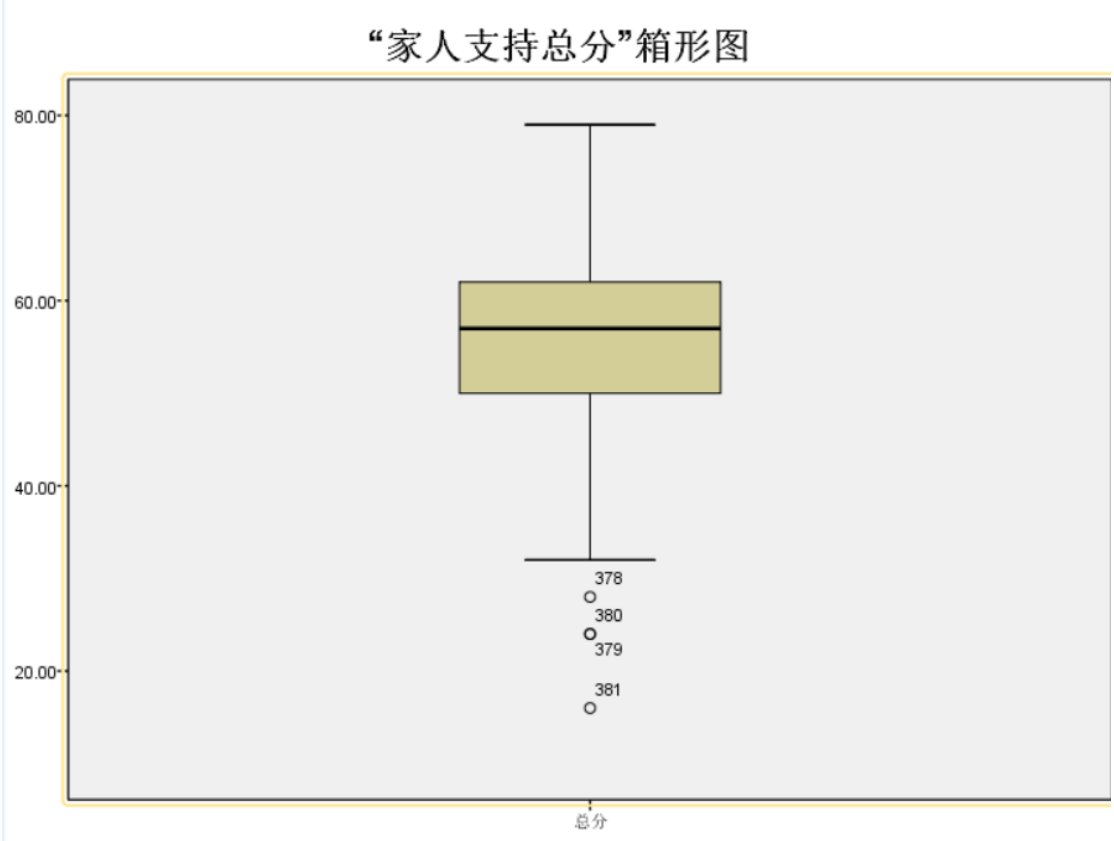
3、在 Q3+1.5IQR（四分位距）和 Q1－1.5IQR 处画两条与中位线一样的线段，这两条线段为异常值截断点，称其为内限；在 Q3+3IQR 和 Q1－3IQR 处画两条线段，称其为外限。处于内限以外位置的点表示的数据都是异常值，其中在内限与外限之间的异常值为温和的异常值（mild outliers），在外限以外的为极端的异常值(extreme outliers)。四分位距 QR=Q3-Q1。.

4、从矩形盒两端边向外各画一条线段直到不是异常值的最远点，表示该批数据正常值的分布区间。

首先，我们对“家人支持”中所有变量绘制箱形图，如下：



图中“o”代表的点为温和异常值，且没有极端异常值的出现。可见，对于取值在 1~5 之间的变量，并没有满足条件的极端异常值。因此，我们进一步对“家人支持总分”绘制箱形图。



可见，仍然只有少数温和的异常值，没有极端异常值。对于极端异常值，我们必须要删去，但是对于温和异常值，应该先分析其数据特征。

7.4.1.2 分析异常值

对于图中的第 381 号变量，我们发现，该答卷者在“家人支持”中的每一项都填上了 1，即“很经常”，可以认为该答卷者没有认真答题，属于无效数据。其他几个温和异常值中，都既有 1，也有 2，因此可以认为是正常数据。

综上，通过箱形图方法检测出的异常值率为 0.26%，说明该数据集可靠性较强。

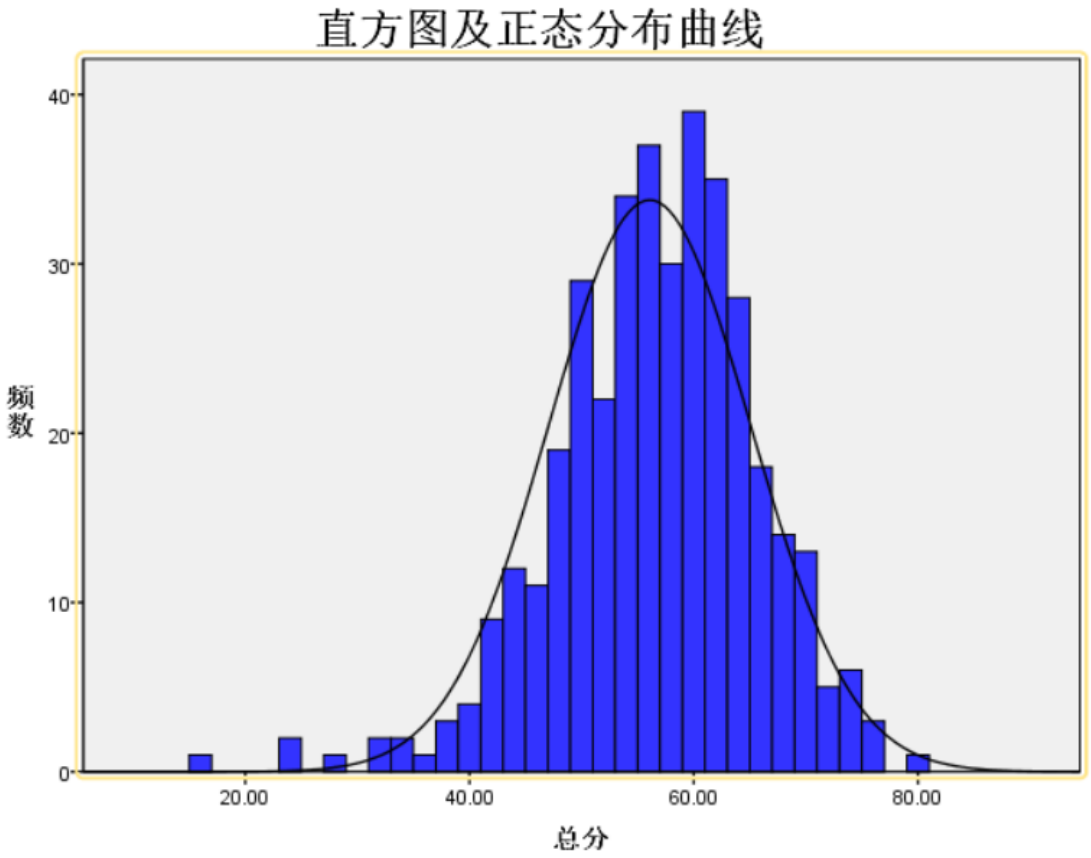
7.4.2 拉依达准则法

7.4.2.1 检验正态分布性

拉依达准则法，对于呈正态分布的变量，筛选距离中心正负 3σ 以外的数据，并定义为离群值，其中 σ 为标准差。根据正态分布理论，变量处于中心正负 3σ

以外的概率为 0.26%。首先，我们检验“家人支持总分”的正态分布性，然后计算离群值在数据集中的频率。若小于 0.26%，则可靠性强。同时需要分析离群值，再决定是否删除。

下图为“家人支持总分”直方图及正态分布曲线：



进一步计算该组数据的峰度系数与偏度系数，分别为 1.380 与-0.612。结合图形，该组数据偏度系数绝对值<1，符合正态分布要求，虽然峰度系数略大于 1，但对于筛选离群值而言，可以放宽对峰度系数的要求，因此我们假定该组数据符合正态分布。

7.4.2.2　筛选离群值

根据上述法则，可以筛选出 4 个离群值，占比 1.05%>0.26%，且这四个离群值恰好为箱形图方法中得出的 4 个异常值。我们将这些离群值删去，再一次计算

数据集的峰度系数与偏度系数，分别为 0.038 和-0.202。可见，删去离群值后，“家人支持总分”较好的符合正太分布，符合实际。

7.4.2.3 分析离群值

由于该方法得出的离群值与箱形图方法得出的异常值相同，因此不做多余的分析。根据箱形图方法中的分析，异常值率为 0.26%，恰好等于正太分布的 3σ 概率，可见，该数据集可靠性较好。

8 结论

逻辑可靠性分析中，除了工作类选项的数据逻辑矛盾率较大以外（23.36%），其他选项均有较好的逻辑一致性。

问卷题项可靠性分析中，可以考虑删去选项 qhc5a 与 qhc6a，提高问卷题项的可靠性。

数据可靠性分析中，异常值\离群值仅有 1 个，占比 0.26%，说明该数据集数据可靠性较好。

9 参考文献

[1]柯惠新，沈浩．调查研究中的统计分析法[M]．中国传媒大学出版社，2005.

[2]吴明隆．问卷统计分析实务[M]．重庆大学出版社，2010.

[3]张士玉．问卷调查数据分析实务[M]．首都经济贸易大学出版社，2015.