中国科学技术大学

《社会科学统计软件与应用：SPSS》结课作业

学生姓名：黄瑞轩

学生学号：PB20111686

学　　院：计算机科学与技术学院

一级学科：计算机科学与技术

指导教师：刘粲

脱贫地区学生基本情况的综合分析

黄瑞轩

（中国科学技术大学　计算机科学与技术学院，计算机科学与技术系）

摘　要

我国脱贫攻坚工作已胜利完成，如何继续针对性的持续帮扶原贫困地区学生，特别是如何让教育脱贫成效稳固可持续，是“后脱贫”时期的重要课题。本研究综合利用 SPSS 和 Python 两种统计分析工具，分析宁夏海原和安徽金寨两处原贫困地区学生样本的调查问卷结果，分析脱贫地区学生基本指标之间的联系，并对今后教育扶贫和“一帮一”活动的工作方向提供建议。

关键词

SPSS；Python；教育扶贫；脱贫；综合分析

1 研究背景与目的

2015 年 11 月 23 日，中共中央政治局审议通过《关于打赢脱贫攻坚战的决定》。中共中央总书记、国家主席、中央军委主席习近平强调，消除贫困、改善民生、逐步实现共同富裕，是社会主义的本质要求，是中国共产党的重要使命。2021 年 2 月 25 日，全国脱贫攻坚总结表彰大会在京隆重举行，习近平庄严宣告：我国脱贫攻坚战取得了全面胜利。

在脱贫攻坚战的伟大实践中，中国科学技术大学芳草社青年志愿者协会（下简称“芳草社”）作为全校志愿者的组织，为学校定点扶贫的宁夏海原县、安徽金寨县做了很多工作。其中影响力最大的是开始于 1997 年、由芳草社主办的“一帮一”启明星导航活动（下简称“一帮一”）。那一年，中国科学技术大学参加“三下乡”的学生将金寨地区孩子的贫寒家境及对上学的渴望这些信息带回了学校，芳草社的志愿者们随即在校内开展了募捐。2004 年底，“一帮一”在宁夏海原县启动。二十多年来，“一帮一”募捐的爱心资助款已达百万元量级，资助学生数以千计。

随着“一帮一”活动的持续开展，芳草社也积累了越来越多的贫困学生资料。这些资料主要分为宁夏海原县二中、三中初中学生信息和安徽金寨县油坊店乡初中学生信息两部分，前者由每年学校派出的研究生支教团带回，后者由每年暑假参加“三下乡”的学生志愿者带回。这些资料以问卷建档的形式保存，一部分是贫困生自己的口述，另一

部分是志愿者对贫困生及其家庭环境的印象陈述。为了在“后脱贫”时期更好地服务这些困难学生、做到“精准”帮助，本研究将对 2019 年至 2021 年三年的贫困生建档问卷进行综合分析。希望分析结果能够更好地帮助芳草社的帮扶志愿工作进行。

2 研究方案与依据

如前文所述，我们的资料是各学生的建档问卷信息，这些问卷都有着标准化的格式，但要用于分析，还需要根据这些问卷信息指定更加详细的量表式问卷。经过对原始问卷的设计分析和对一些学者发表的致贫原因分析报告的综合，本研究将考虑如下因素来设计详细问卷。

（1）人口学特征：生活地区、性别、年龄、家庭结构、共同生活人数、劳动能力等方面。生活地区根据调查编号即可得出（海原二中编号以 EZ 开头、海原三中编号以 SZ 开头、金寨油坊店乡编号以 JZ 开头）；家庭结构可选择的选项有双亲健在、父母离婚（母亲抚养）、父母离婚（父亲抚养）、父亲逝世（母亲抚养）、母亲去世（父亲抚养）和双亲逝世（其他亲属抚养）；劳动能力指的是学生主要抚养人的劳动力情况，依照《国务院扶贫办关于做好 2019 年度扶贫对象动态管理工作的通知》（国开办发〔2019〕14 号）中的指标解释。

（2）患病情况：学生本人和家庭成员的疾病数量、疾病分类等方面。疾病数量依照资料叙述如实登记，疾病分类分为大病、慢病和常见多发病三种。其中，大病定义为医治花费巨大且在较长一段时间内严重影响患者及其家庭的正常工作和生活的疾病，一般包括：恶性肿瘤、严重心脑血管疾病；慢病定义为无需手术或者手术后需要长期服药控制病情、但是患者可以参与部分生产劳动的疾病，一般包括：风湿病、肌肉劳损等；常见多发病则指一些于患者中常见的疾病，如糖尿病、高血压、普通型心脏病等。

（3）学习情况：学习阶段、学习情况等方面。调查对象的学习阶段绝大多数都是初中一年级（即七年级）。针对志愿者、老师对调查对象学习情况的印象，按照五点法分类为很差、较差、中等、较好、优异五个水平。

（4）收入和支配情况：家庭收入来源等方面。根据家庭主要劳动力获得收入的方式的不同分类此项，包括务农、打工、亲戚资助等。

处理已有数据的主要工具是 Python。首先，根据编号前缀的不同将学生分类，分类变量为Location。由于尚不知年份对结果的影响，还需要通过编号新增一个调查年份 Year 变量。家庭结构按照双亲健在、父母离婚（母亲抚养）、父母离婚（父亲抚养）、父亲逝世（母亲抚养）、母亲去世（父亲抚养）和双亲逝世（其他亲属抚养）的顺序分别定义为 0～5，按照格式分成父亲正抚养、母亲正抚养、父亲正健在、母亲正健在的四个变

量，其中前二者是布尔变量，后二者有三值：是、否和不清楚。劳动能力按照三点法评价。患病情况中，本人和家人的患病种类按照三点法评价，从相对最严重到相对最不严重分别为大病、慢病和常见多发病。学习情况按照李特克五点法评价，从相对最差到相对最好。家庭收入来源则是所有情况的开关变量，包含务农、打工、亲戚资助、拾荒、低保等。将原来的 xlsx 格式数据转换为 csv 格式数据后，利用 Python 的 csv 库可直接将数据转换为可供 SPSS 读取分析的数据。

分析数据的主要工具是 SPSS。由于海原和金寨相隔距离较远、各方面情况迥异，因此认为海原和金寨的样本间是互相独立的。研究过程中，我们将综合应用 SPSS 的功能，对两地困难学生的信息关联做统计学上的分析。

3 研究过程

3.1 综合因素概览

我们首先从整体的角度来看一下各重要因子的分布。2019 年参与的地区为海原二中、金寨油坊店；2020 年参与的地区为海原二中、海原三中；2021 年参与的地区为海原二中、海原三中、金寨油坊店。人数分布如图 1 所示。

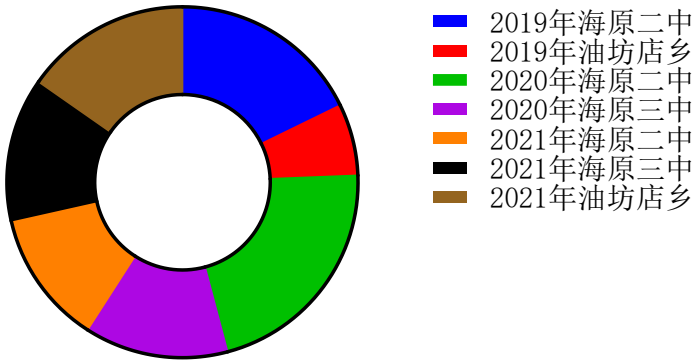

图 1 各年份各地区人数分布

由于初中二年级同学只有 2020 年金寨地区的 4 名同学，这里暂不考虑。各初中一年级同学参与登记时的年龄分布如图 2 所示。其中年龄最小的同学 11 岁，年龄最大的同学 17 岁。占比最大的人群为 13 岁，占总人数的 34.3%。从全国来看，正常升入初一的学生年龄在 12～14 岁间。在本样本中，12～14 岁人群占比 81.7%，说明这些地区孩子的上学时期并没有在较大程度上受到贫困条件的影响。

图 2 各初一同学参与登记时的年龄分布

从性别来看，女性占 65.7%，男性占 34.3%。女性占几乎 2/3，这是值得研究的问题。

就整体的家庭结构来看，父母离婚的孩子占比 19.2%；父母有一方去世或离开孩子的占比 14.2%；父母双方均不在世或均离开孩子的占比 10.0%。根据国家统计局数据，2020 年我国粗离婚率为 3.09‰[1]，2019 年我国留守儿童占比 6.4%[2]。可以粗略感受到在这些地区，不正常的家庭结构占比显著偏大，父母一方或者双方的缺失在很多方面都会对孩子的发展产生不利影响。

而细分来看，双亲、单独父亲、单独母亲和其他亲属的抚养占比如图 3 所示。在所有样本中，由双亲抚养的情况最多（占比 56.5%），但仍然有接近一半的孩子由单亲或其他亲属抚养，这之中又以母亲单独抚养的比例为最多。

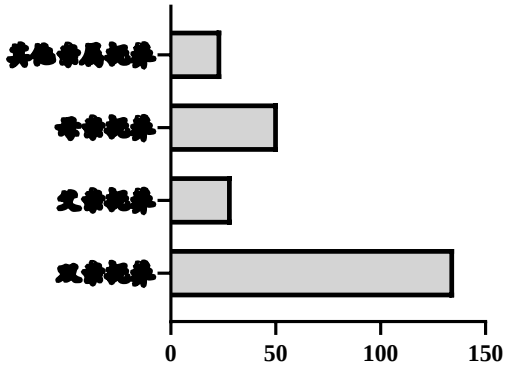

图 3 各抚养情况占比

从共同生活的人口数来看，各地区分别统计的区间情况如表 1 所示。统计时，以小于 4 人、4 至 6 人和大于 6 人为分界。通过统计结果我们发现在所有地区，4 至 6 人的情况都是最多的；大于 6 人的情况基本没有。处理数据时也发现，当父母都健在时，孩子往往不填写更年长的家人的信息。绝大部分孩子填写的信息只包括同辈和父母辈的家人的信息，具体更年长的家人是否与之生活在一起则不得而知，这里假定孩子们填写的信息就是他们所处的共同生活环境的人口数。各地区分别统计的详细情况如图 4 所示。

	小于 4 人	4 至 6 人	大于 6 人
金寨油坊店	19	34	0
海原二中	37	87	1
海原三中	21	39	1

表 1 各地区分别统计的区间情况

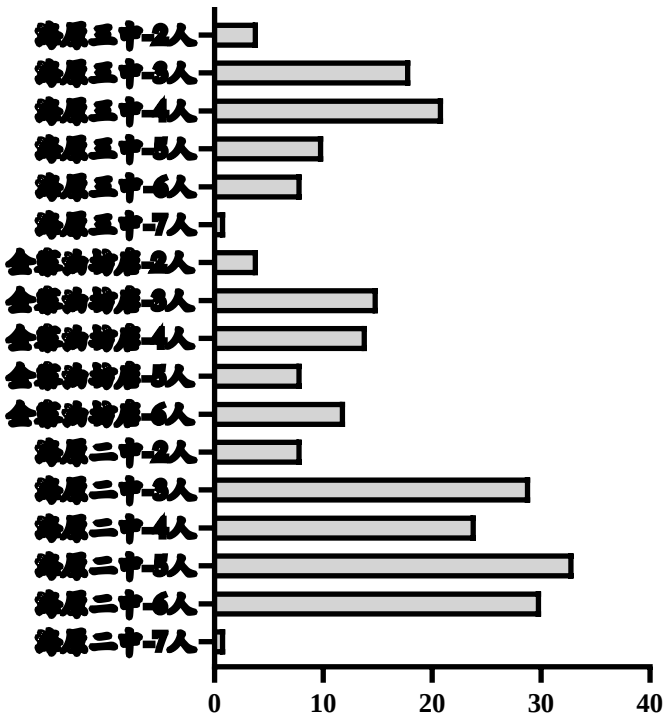

图 4 各地区分别统计的详细情况

从家庭主要劳动力情况来看，普通劳动力计数 196，占比 82.0%；弱劳动力计数 32，占比 13.4%；无劳动力计数 11，占比 4.6%。

从孩子本人的患病情况来看，在 239 个个案中只有 3 个是孩子本人患病的，均为慢病。从家庭成员患病情况来看，患三种病、患两种病、患一种病和不患病的比例分别是

0.4%、10.9%、31.0%和 57.7%。患病人群中，患常见多发病、慢病和大病的比例分别是 49.5、26.7 和 23.8%。有接近一半的个案是家庭成员有患病的。在患病人群中，有超过一半是患慢病和大病的。

从孩子的学习成绩看，有 99 个个案未提供相关信息，这里仅讨论有信息的 140 个个案。这之中，学习成绩从很差到优异的五个阶段个案占比分别为 0、17.1%、32.9%、12.9%和 37.1%。成绩中等及以上的同学占了绝大多数。

从家庭收入来源方式看，共有打工、务农、亲戚资助、低保和拾荒五种来源，分别占比 28.5%、55.6%、9.6%、13.0%和 0.4%。务农和打工是占绝大多数的收入来源。这里同一个家庭可能有多种收入来源，所以总的百分比可能大于 100%。

3.2 家庭结构对学习成绩影响分析

研究表明，离婚主要发生在经济较困难的家庭，而离婚会进一步导致家庭经济收入的下降[3]。家庭变故导致一方亲人逝世也会对孩子产生不利的影响。这一部分研究家庭结构对孩子学习成绩的影响。家庭结构在本研究中体现为抚养情况，具体分类为父亲抚养 F 和母亲抚养 M 两个布尔变量。学习成绩作为被解释变量，定义为 Re。由于有一部分孩子没有提供学习成绩相关数据，这里的分析是建立在提供了学习成绩的孩子样本上的。我们假设 Re 与诸因素之间存在线性相关关系，因此做出如下三种假定：

$$(\text{Ⅰ})\ Re = \alpha(F + M) + \beta S + \varepsilon$$

$$(\text{Ⅱ})\ Re = \alpha(F + M) + \varepsilon$$

$$(\text{Ⅲ})\ Re = \beta S + \varepsilon$$

其中 α、β 是待定系数，S 是孩子性别（考虑到对男孩和女孩的投入可能不同，这里将女孩的 S 变量赋值为 0，将男孩的 S 变量赋值为 1），ε 是随机误差。在 SPSS 中实际计算时，F + M 的名字为 Foster，S 的名字为 Sex。

对这三种假定，分别采用 SPSS 的多变量线性回归分析，分析结果如结果 1、结果 2 和结果 3 所示。

模型摘要

模型	R	R 方	调整后 R 方	标准估算的错误
1	.214[a]	.046	.032	1.124

a. 预测变量：(常量), Sex, Foster

ANOVA[a]

模型		平方和	自由度	均方	F	显著性
1	回归	8.331	2	4.166	3.297	.040[b]
	残差	173.069	137	1.263		
	总计	181.400	139			

a. 因变量：Record

b. 预测变量：(常量), Sex, Foster

系数[a]

模型		未标准化系数		标准化系数		
		B	标准错误	Beta	t	显著性
1	(常量)	3.059	.269		11.354	.000
	Foster	.363	.157	.194	2.315	.022
	Sex	.169	.196	.072	.862	.390

a. 因变量：Record

结果 1 (Ⅰ)模型的回归分析结果

模型摘要

模型	R	R 方	调整后 R 方	标准估算的错误
1	.202[a]	.041	.034	1.123

a. 预测变量：(常量), Foster

ANOVA[a]

模型		平方和	自由度	均方	F	显著性
1	回归	7.392	1	7.392	5.862	.017[b]
	残差	174.008	138	1.261		
	总计	181.400	139			

a. 因变量：Record

b. 预测变量：(常量), Foster

系数[a]

模型		未标准化系数		标准化系数		
		B	标准错误	Beta	t	显著性
1	(常量)	3.102	.265		11.723	.000
	Foster	.377	.156	.202	2.421	.017

a. 因变量：Record

结果 2　(Ⅱ)模型的回归分析结果

模型摘要

模型	R	R 方	调整后 R 方	标准估算的错误
1	.093[a]	.009	.001	1.142

a. 预测变量：(常量), Sex

ANOVA[a]

模型		平方和	自由度	均方	F	显著性
1	回归	1.563	1	1.563	1.199	.275[b]
	残差	179.837	138	1.303		
	总计	181.400	139			

a. 因变量：Record

b. 预测变量：(常量), Sex

系数[a]

模型		未标准化系数		标准化系数		
		B	标准错误	Beta	t	显著性
1	(常量)	3.616	.123		29.377	.000
	Sex	.217	.198	.093	1.095	.275

a. 因变量：Record

结果 3　(III)模型的回归分析结果

在结果 1 中，模型摘要体现了模型的拟合程度。调整后 R 方为 0.032，拟合度为 3.2%，拟合程度较差。ANOVA 体现了整体的显著性。这里 *p*-value 为 0.04，小于 0.05，说明整体显著，至少有一个自变量对因变量有显著性影响。系数是回归系数的结果。这里 Sex 和 Foster 的回归系数均为正，说明都对 Record 有正向影响。其中 Sex 的 *p*-value 大于 0.05，影响不显著，Foster 的 *p*-value 小于 0.05，有显著性影响。

以上分析结果表明，虽然在大部分人的印象中贫困地区可能存在着重男轻女的思想，因此对男孩的教育投入可能大于女孩，进而在性别方面影响分析结果。但分析结果指出，Sex 变量对 Record 的影响是不显著的，也就是说性别对学习成绩的影响是不显著的。但双亲的抚养对学习成绩的影响是显著的，并且是正向的影响。这说明在家庭结构正常的个案中，孩子的学习成绩要更好一些；而在家庭结构不正常——诸如遭遇变故去世或离婚——的家庭中，孩子的学习成绩更倾向于在一个较低的水平。

3.3 家庭劳动力及收入来源对学习成绩影响分析

家庭劳动力是家庭收入的关键因素，家庭收入在很大程度上影响家庭对孩子包括教育在内的各方面的投入。家庭的收入来源方式可能也会对孩子产生一些影响。这一部分主要分析的是家庭劳动力水平及家庭收入来源对孩子学习成绩的影响。由于金寨和海原地区不同，家庭劳动力及收入来源方式可能之间存在差异，因此不适用于整体研究，这一部分将分区域分析。

家庭劳动力在本研究中已经按三点法分类。收入来源已被细分成五个布尔变量，这些收入来源方式由于稳定性的不同，不宜直接相加作为自变量。因此在本研究中，每种收入来源方式将作为一个单独的自变量。打工、务农、亲戚资助、低保和拾荒五种收入来源方式分别记为 X_1、X_2、X_3、X_4 和 X_5，家庭劳动力记为 C，由此建立的假设模型为：

$$\mathrm{Re} = x\mathrm{C} + y\mathrm{X}_1 + z\mathrm{X}_2 + u\mathrm{X}_3 + v\mathrm{X}_4 + w\mathrm{X}_5$$

其中 x、y、z、u、v 和 w 都是待定系数。在 SPSS 实际计算中，C 的名字是 Capacity，X_1～X_5 的名字分别是 Working、Farming、Funded、Allowance 和 Scavenging。对这个假设模型采用多变量回归分析，结果如结果 4 所示。

从结果 4 来看这一假设是不合适的，因为所有的 p-value 均大于 0.05，结果均不显著。这表明从我们有限的样本中，并不能得出家庭劳动力与家庭收入来源对孩子学习成绩有影响的结论。

模型摘要

模型	R	R 方	调整后 R 方	标准估算的错误
1	.161[a]	.026	-.018	1.153

a. 预测变量：(常量), Allowance, Scavenging, Funded, Farming, Capacity, Working

ANOVA[a]

模型		平方和	自由度	均方	F	显著性
1	回归	4.724	6	.787	.593	.736[b]
	残差	176.676	133	1.328		
	总计	181.400	139			

a. 因变量：Record

b. 预测变量：(常量), Allowance, Scavenging, Funded, Farming, Capacity, Working

系数[a]

		未标准化系数		标准化系数		
模型		B	标准错误	Beta	t	显著性
1	(常量)	4.140	.785		5.272	.000
	Capacity	-.049	.227	-.021	-.217	.829
	Farming	-.071	.385	-.028	-.185	.853
	Working	-.319	.427	-.137	-.747	.456
	Funded	-.396	.548	-.089	-.721	.472
	Scavenging	-1.971	1.189	-.146	-1.658	.100
	Allowance	-.408	.550	-.114	-.742	.460

a. 因变量：Record

结果 4　C 和 X_1～X_5 的回归分析

3.3 地区差异性检验

在整体研究之后，我们自然要问：不同地区（金寨、海原）对各学生信息指标是否有影响？我们可以采用 SPSS 的方差分析功能，所采用的分类变量为 Location，因变量是除了 Year（年份）和 Grade（学习阶段）之外所有的变量。得到结果如结果 5 所示。

ANOVA

		平方和	自由度	均方	F	显著性
Sex	组间	.083	2	.041	.181	.834
	组内	53.783	236	.228		
	总计	53.866	238			
Age	组间	71.302	2	35.651	36.349	.000
	组内	231.468	236	.981		
	总计	302.770	238			
Paternity	组间	.732	2	.366	1.703	.184
	组内	50.732	236	.215		
	总计	51.464	238			
Motherhood	组间	1.096	2	.548	3.222	.042
	组内	40.150	236	.170		
	总计	41.247	238			
FatherLiving	组间	1.646	2	.823	1.172	.312
	组内	165.726	236	.702		
	总计	167.372	238			
MotherLiving	组间	2.582	2	1.291	2.270	.106
	组内	134.229	236	.569		
	总计	136.812	238			
OtherRalatives	组间	.471	2	.236	.647	.524
	组内	85.889	236	.364		
	总计	86.360	238			
NumofLivingTogether	组间	5.068	2	2.534	1.624	.199
	组内	368.254	236	1.560		
	总计	373.322	238			
Capacity	组间	.053	2	.026	.098	.907
	组内	63.746	236	.270		
	总计	63.799	238			
NumofDiseaseHimself	组间	.013	2	.007	.529	.590

	组内	2.949	236	.012		
	总计	2.962	238			
TypeofDiseaseHimself	组间	.053	2	.026	.529	.590
	组内	11.797	236	.050		
	总计	11.849	238			
NumofDiseaseHisFamily	组间	.314	2	.157	.316	.729
	组内	117.059	236	.496		
	总计	117.372	238			
TypeofDiseaseHisFamily	组间	.837	2	.418	.406	.667
	组内	243.556	236	1.032		
	总计	244.393	238			
Record	组间	4.740	2	2.370	1.838	.163
	组内	176.660	137	1.289		
	总计	181.400	139			
Farming	组间	.129	2	.065	.314	.731
	组内	48.524	236	.206		
	总计	48.653	238			
Working	组间	.059	2	.029	.118	.889
	组内	58.929	236	.250		
	总计	58.987	238			
Funded	组间	.031	2	.015	.174	.840
	组内	20.756	236	.088		
	总计	20.787	238			
Allowance	组间	.082	2	.041	.362	.697
	组内	26.897	236	.114		
	总计	26.979	238			
Scavenging	组间	.012	2	.006	1.465	.233
	组内	.984	236	.004		
	总计	.996	238			

结果 5　按 Location 分类的 ANOVA 检验

从方差分析的 *p*-value 来看，只有 Motherhood 的 *p*-value 小于 0.05，说明在这两个不同的地区，母亲的抚养情况存在显著差异。其余因变量的 p-value 都大于 0.05，因此在这两个不同地区其他指标的差异是不显著的。

3.4 时间差异性检验

我们收集的数据来自 2019 至 2021 三年，这之中的 2020 年中国遭遇了新冠肺炎疫情，经济发展放缓。我们自然也关心我们的数据是否有时间上的差异性？我们采用 SPSS 的方差分析功能，所采用的分类变量为 Year，因变量是除了 Location（地区）和 Grade（学习阶段）之外所有的变量。得到结果如结果 6 所示。

ANOVA

		平方和	自由度	均方	F	显著性
Sex	组间	.616	2	.308	1.364	.258
	组内	53.251	236	.226		
	总计	53.866	238			
Age	组间	31.686	2	15.843	13.792	.000
	组内	271.084	236	1.149		
	总计	302.770	238			
Paternity	组间	.607	2	.304	1.408	.247
	组内	50.857	236	.215		
	总计	51.464	238			
Motherhood	组间	2.109	2	1.055	6.360	.002
	组内	39.137	236	.166		
	总计	41.247	238			
FatherLiving	组间	.990	2	.495	.702	.496
	组内	166.382	236	.705		
	总计	167.372	238			
MotherLiving	组间	7.234	2	3.617	6.587	.002
	组内	129.578	236	.549		
	总计	136.812	238			
OtherRalatives	组间	1.936	2	.968	2.706	.069
	组内	84.424	236	.358		
	总计	86.360	238			
NumofLivingTogether	组间	1.798	2	.899	.571	.566
	组内	371.524	236	1.574		
	总计	373.322	238			
Capacity	组间	.447	2	.223	.833	.436
	组内	63.352	236	.268		
	总计	63.799	238			
NumofDiseaseHimself	组间	.040	2	.020	1.625	.199

	组内	2.922	236	.012		
	总计	2.962	238			
TypeofDiseaseHimself	组间	.161	2	.080	1.625	.199
	组内	11.688	236	.050		
	总计	11.849	238			
NumofDiseaseHisFamily	组间	.158	2	.079	.159	.853
	组内	117.214	236	.497		
	总计	117.372	238			
TypeofDiseaseHisFamily	组间	.197	2	.099	.095	.909
	组内	244.196	236	1.035		
	总计	244.393	238			
Record	组间	.214	1	.214	.163	.687
	组内	181.186	138	1.313		
	总计	181.400	139			
Farming	组间	.072	2	.036	.176	.839
	组内	48.580	236	.206		
	总计	48.653	238			
Working	组间	.908	2	.454	1.845	.160
	组内	58.079	236	.246		
	总计	58.987	238			
Funded	组间	.226	2	.113	1.297	.275
	组内	20.561	236	.087		
	总计	20.787	238			
Allowance	组间	.082	2	.041	.361	.697
	组内	26.897	236	.114		
	总计	26.979	238			
Scavenging	组间	.008	2	.004	.975	.379
	组内	.988	236	.004		
	总计	.996	238			

结果 6　按 Year 分类的 ANOVA 检验

从方差分析的 *p*-value 结果看，Motherhood（母亲抚养情况）和 Mother Living（母亲存活情况）的 *p*-value 均小于 0.05，时间上的差异性是显著的；其他因变量的 *p*-value 都大于 0.05，因此在时间上的差异性是不显著的。

3.5 性别差异性检验

如 3.1 节中所述，我们的样本指标概览指出，接近 2/3 的个案为女性，仅约 1/3 的个案为男性。有学者提出性别差异可能会影响人的自我认同[4]。这样的影响可能会对包括学习成绩在内的一些指标产生影响。性别差异是否会引起其他指标的差异？我们采用 SPSS 的方差分析功能，所采用的分类变量为 Sex，因变量是除了 Location（地区）、Year（年份）和 Grade（学习阶段）之外所有的变量。得到结果如结果 7 所示。

ANOVA

		平方和	自由度	均方	F	显著性
Age	组间	.632	1	.632	.496	.482
	组内	302.137	237	1.275		
	总计	302.770	238			
Paternity	组间	.610	1	.610	2.843	.093
	组内	50.854	237	.215		
	总计	51.464	238			
Motherhood	组间	.188	1	.188	1.086	.298
	组内	41.059	237	.173		
	总计	41.247	238			
FatherLiving	组间	2.354	1	2.354	3.380	.067
	组内	165.019	237	.696		
	总计	167.372	238			
MotherLiving	组间	1.161	1	1.161	2.028	.156
	组内	135.651	237	.572		
	总计	136.812	238			
OtherRalatives	组间	.777	1	.777	2.151	.144
	组内	85.583	237	.361		
	总计	86.360	238			
NumofLivingTogether	组间	3.285	1	3.285	2.104	.148
	组内	370.038	237	1.561		
	总计	373.322	238			
Capacity	组间	.040	1	.040	.150	.699
	组内	63.759	237	.269		
	总计	63.799	238			
NumofDiseaseHimself	组间	.017	1	.017	1.408	.237
	组内	2.945	237	.012		

	总计	2.962	238			
TypeofDiseaseHimself	组间	.070	1	.070	1.408	.237
	组内	11.779	237	.050		
	总计	11.849	238			
NumofDiseaseHisFamily	组间	.029	1	.029	.059	.808
	组内	117.343	237	.495		
	总计	117.372	238			
TypeofDiseaseHisFamily	组间	1.716	1	1.716	1.676	.197
	组内	242.677	237	1.024		
	总计	244.393	238			
Record	组间	1.563	1	1.563	1.199	.275
	组内	179.837	138	1.303		
	总计	181.400	139			
Farming	组间	.405	1	.405	1.988	.160
	组内	48.248	237	.204		
	总计	48.653	238			
Working	组间	.245	1	.245	.988	.321
	组内	58.743	237	.248		
	总计	58.987	238			
Funded	组间	.155	1	.155	1.783	.183
	组内	20.631	237	.087		
	总计	20.787	238			
Allowance	组间	.354	1	.354	3.147	.077
	组内	26.626	237	.112		
	总计	26.979	238			
Scavenging	组间	.002	1	.002	.521	.471
	组内	.994	237	.004		
	总计	.996	238			

结果 7　按 Sex 分类的 ANOVA 检验

检验结果指出，包括 Record（学习成绩）在内的所有指标的 *p*-value 均大于 0.05。这说明我们收集到的指标中，没有在性别差异上产生显著差异的指标。

3.6 家庭患病情况差异性检验

据研究，因病致贫患者共病比例和疾病严重程度均明显高于非因病致贫患者，重大疾病、慢性病是农村贫困地区致贫、返贫的主要原因[5]。家庭患病情况的差异会对其他指标产生差异吗？我们定义的家庭患病情况如下：

$$H = N_1 * S_1 + N_2 * S_2$$

其中，H 代表家庭患病情况，N_1 与 N_2 分别代表个人患病数和家庭患病数，S_1 和 S_2 分别代表个人患病严重度和家庭患病严重度。这里的严重度可以用按三点法评价的疾病分类来估计，但由于数据中数字越小疾病越严重，如果用 s_1 和 s_2 表示个人患病分类和家庭成员患病分类的数值，我们在检验时要做如下处理：

$$S_x = \max\{s_i\} + 1 - s_x \ (x = 1, 2)$$

即计算三个新的变量，名字记为 S1、S2 和 H。我们采用 SPSS 的方差分析功能，所采用的分类变量为 H，因变量是除了 Location（地区）、Year（年份）、Grade（学习阶段）以及与 H 的构造有关的变量之外所有的变量。得到结果如结果 8 所示。

ANOVA

		平方和	自由度	均方	F	显著性
Age	组间	14.111	6	2.352	1.890	.083
	组内	288.658	232	1.244		
	总计	302.770	238			
Sex	组间	1.453	6	.242	1.072	.380
	组内	52.413	232	.226		
	总计	53.866	238			
Paternity	组间	.840	6	.140	.642	.697
	组内	50.624	232	.218		
	总计	51.464	238			
Motherhood	组间	.513	6	.086	.487	.818
	组内	40.734	232	.176		
	总计	41.247	238			
FatherLiving	组间	5.698	6	.950	1.363	.231
	组内	161.675	232	.697		
	总计	167.372	238			
MotherLiving	组间	1.491	6	.248	.426	.861
	组内	135.321	232	.583		
	总计	136.812	238			
OtherRalatives	组间	1.211	6	.202	.550	.770

	组内	85.149	232	.367		
	总计	86.360	238			
NumofLivingTogether	组间	7.752	6	1.292	.820	.555
	组内	365.570	232	1.576		
	总计	373.322	238			
Capacity	组间	29.084	6	4.847	32.394	.000
	组内	34.715	232	.150		
	总计	63.799	238			
Record	组间	9.723	6	1.620	1.255	.282
	组内	171.678	133	1.291		
	总计	181.400	139			
Farming	组间	1.777	6	.296	1.466	.191
	组内	46.875	232	.202		
	总计	48.653	238			
Working	组间	2.867	6	.478	1.976	.070
	组内	56.120	232	.242		
	总计	58.987	238			
Funded	组间	.879	6	.146	1.707	.120
	组内	19.908	232	.086		
	总计	20.787	238			
Allowance	组间	2.774	6	.462	4.431	.000
	组内	24.206	232	.104		
	总计	26.979	238			
Scavenging	组间	.062	6	.010	2.589	.019
	组内	.933	232	.004		
	总计	.996	238			

结果 8　按 H 分类的 ANOVA 检验

从结果 8 的 *p*-value 情况中我们可以看到，Allowance（低保）和 Scavenging（拾荒）的 *p*-value 都小于 0.05，是差异显著的。这说明在患病情况不同时，低保的接受情况有显著差异；拾荒的出现情况也有显著差异。

值得注意的是，我们定义的 H 将个人患病情况与家庭成员患病情况混为一谈。如果更关心家庭成员的患病情况（劳动收入的主要影响因素），即命：

$$H' = S_2$$

再次进行检验，即可获得如结果 9 所示的结果。

ANOVA

		平方和	自由度	均方	F	显著性
Age	组间	3.952	3	1.317	1.036	.377
	组内	298.818	235	1.272		
	总计	302.770	238			
Sex	组间	.660	3	.220	.972	.407
	组内	53.206	235	.226		
	总计	53.866	238			
Paternity	组间	.270	3	.090	.413	.744
	组内	51.194	235	.218		
	总计	51.464	238			
Motherhood	组间	.203	3	.068	.387	.763
	组内	41.044	235	.175		
	总计	41.247	238			
FatherLiving	组间	2.188	3	.729	1.038	.377
	组内	165.184	235	.703		
	总计	167.372	238			
MotherLiving	组间	.461	3	.154	.265	.851
	组内	136.351	235	.580		
	总计	136.812	238			
OtherRalatives	组间	.500	3	.167	.456	.713
	组内	85.860	235	.365		
	总计	86.360	238			
NumofLivingTogether	组间	.323	3	.108	.068	.977
	组内	372.999	235	1.587		
	总计	373.322	238			
Capacity	组间	25.164	3	8.388	51.022	.000
	组内	38.635	235	.164		
	总计	63.799	238			
Record	组间	2.752	3	.917	.698	.555
	组内	178.648	136	1.314		
	总计	181.400	139			
Farming	组间	.713	3	.238	1.165	.324
	组内	47.939	235	.204		

	总计	48.653	238			
Working	组间	2.233	3	.744	3.082	.028
	组内	56.754	235	.242		
	总计	58.987	238			
Funded	组间	.374	3	.125	1.434	.234
	组内	20.413	235	.087		
	总计	20.787	238			
Allowance	组间	1.690	3	.563	5.233	.002
	组内	25.290	235	.108		
	总计	26.979	238			
Scavenging	组间	.037	3	.012	3.064	.029
	组内	.958	235	.004		
	总计	.996	238			

结果 9　按 S2 分类的 ANOVA 检验

结果 9 指出，除了 Allowance 和 Scavenging，还有 Working（打工）情况也是 *p*-value 小于 0.05 的。这说明在 H′的意义下，也可以说患病情况不同时，以打工为主要劳动收入来源的情况也具有显著的差异。

3.7 家庭共同生活人口数差异性检验

如 3.1 节中所指出的，在所有地区个案中 4～6 人生活在一起的情况占比最大。我们知道在贫困条件下想要同时养活较多的人口数是比较困难的。家庭共同生活人口数的差异会对包括 Record（学习成绩）在内的各指标产生差异性影响吗？我们采用 SPSS 的方差分析功能，所采用的分类变量为 NLT（NumofLivingTogether），因变量是除了 Location（地区）、Year（年份）和 Grade（学习阶段）之外所有的变量。得到结果如结果 10 所示。

ANOVA

		平方和	自由度	均方	F	显著性
Age	组间	5.929	5	1.186	.931	.462
	组内	296.841	233	1.274		
	总计	302.770	238			
Sex	组间	1.833	5	.367	1.642	.150
	组内	52.033	233	.223		
	总计	53.866	238			
Paternity	组间	5.042	5	1.008	5.061	.000
	组内	46.423	233	.199		
	总计	51.464	238			
Motherhood	组间	5.520	5	1.104	7.200	.000
	组内	35.727	233	.153		
	总计	41.247	238			
FatherLiving	组间	17.066	5	3.413	5.291	.000
	组内	150.307	233	.645		
	总计	167.372	238			
MotherLiving	组间	18.330	5	3.666	7.210	.000
	组内	118.481	233	.509		
	总计	136.812	238			
OtherRalatives	组间	15.018	5	3.004	9.810	.000
	组内	71.342	233	.306		
	总计	86.360	238			
Capacity	组间	1.250	5	.250	.931	.461
	组内	62.549	233	.268		
	总计	63.799	238			
NumofDiseaseHimself	组间	.066	5	.013	1.061	.383
	组内	2.896	233	.012		

	总计	2.962	238			
TypeofDiseaseHimself	组间	.264	5	.053	1.061	.383
	组内	11.586	233	.050		
	总计	11.849	238			
NumofDiseaseHisFamily	组间	1.806	5	.361	.728	.603
	组内	115.567	233	.496		
	总计	117.372	238			
TypeofDiseaseHisFamily	组间	3.792	5	.758	.734	.598
	组内	240.601	233	1.033		
	总计	244.393	238			
Record	组间	4.663	5	.933	.707	.619
	组内	176.737	134	1.319		
	总计	181.400	139			
Farming	组间	.979	5	.196	.957	.445
	组内	47.674	233	.205		
	总计	48.653	238			
Working	组间	1.024	5	.205	.823	.534
	组内	57.964	233	.249		
	总计	58.987	238			
Funded	组间	.775	5	.155	1.804	.113
	组内	20.012	233	.086		
	总计	20.787	238			
Allowance	组间	.107	5	.021	.186	.968
	组内	26.872	233	.115		
	总计	26.979	238			
Scavenging	组间	.062	5	.012	3.120	.010
	组内	.933	233	.004		
	总计	.996	238			

结果 10　按 NLT 分类的 ANOVA 检验

结果 10 指出，NLT 的差异在 ANOVA 检验下显著的指标主要是抚养情况。进一步地，我们假定一个回归模型：

$$NLT = xY_1 + yY_2 + zY_3$$

其中 x、y 和 z 都是待定系数。在 SPSS 实际计算中，Y_1～Y_3 是父母抚养情况以及其他亲属抚养情况。对这个假设模型采用多变量回归分析，结果如结果 11 所示。

模型摘要[b]

模型	R	R 方	调整后 R 方	标准估算的错误
1	.381[a]	.145	.134	1.166

a. 预测变量：(常量), OtherRalatives, Paternity, Motherhood

b. 因变量：NumofLivingTogether

ANOVA[a]

模型		平方和	自由度	均方	F	显著性
1	回归	54.084	3	18.028	13.271	.000[b]
	残差	319.238	235	1.358		
	总计	373.322	238			

a. 因变量：NumofLivingTogether

b. 预测变量：(常量), OtherRalatives, Paternity, Motherhood

系数[a]

模型		未标准化系数		标准化系数		
		B	标准错误	Beta	t	显著性
1	(常量)	3.466	.187		18.519	.000
	Paternity	.018	.192	.007	.096	.924
	Motherhood	.454	.239	.151	1.905	.058
	OtherRalatives	-.550	.187	-.264	-2.937	.004

a. 因变量：NumofLivingTogether

结果 11　诸亲人抚养因素对 NLT 的回归分析

结果 11 指出，从 *p*-value 的角度，NLT 主要受 Motherhood（母亲抚养）和 OtherRalatives（其他亲属抚养）的影响。其中，Paternity 和 Motherhood 的回归系数是正的，说明有双亲或者其一方抚养的家庭共同生活人口数更多；OtherRalatives 的回归系数是负的，符合统计规则，说明仅有孩子和抚养其的亲人，所以家庭共同生活人口数较少。

4 研究信息汇总

经过如上的统计分析，我们从已掌握的数据中得到了如下分析结果。

（1）样本各地区孩子的上学时期并没有在较大程度上受到贫困条件的影响；

（2）在样本各地区，不正常的家庭结构占比显著偏大；

（3）在所有样本中，由双亲抚养的情况最多，但仍然有接近一半的孩子由单亲或其他亲属抚养，这之中又以母亲单独抚养的比例为最多；

（4）有接近一半的个案是家庭成员有患病的。在患病人群中，有超过一半是患慢病和大病的；

（5）性别对学习成绩的影响是不显著的，但双亲的抚养对学习成绩的影响是显著的，并且是正向的影响；

（6）家庭劳动力与家庭收入来源方式对孩子学习成绩的影响是不显著的；

（7）在不同地区，母亲的抚养情况存在显著差异，其他则没有显著差异；

（8）在不同年份，母亲抚养情况和母亲存活情况存在显著差异，其他则没有显著差异；

（9）我们收集到的指标中，没有在性别差异上产生显著差异的；

（10）在患病情况不同时，低保的接受情况有显著差异，拾荒的出现情况也有显著差异；修正后，也可以说患病情况不同时，以打工为主要劳动收入来源的情况也具有显著的差异；

（11）共同生活的人口数（NLT）的差异在 ANOVA 检验下显著的指标主要是抚养情况，NLT 主要受母亲抚养和其他亲属抚养的影响。

5 研究总结与建议

经过数据处理与分析，我们得出了一些关于样本数据的有用的结论。这些结论清扫了我们关于贫困孩子困难因素的一些刻板印象，其结果有助于我们增强“一帮一”活动帮扶的针对性。下面列出几条从研究结论出发，对“一帮一”活动有指导意义的建议。

建议 1：在给学校志愿者或愿意帮扶的同学提供信息时，不需要过多的考虑年龄的因素，按照一般初中学生的水平提供帮扶即可。

建议 2：在提供书信往来志愿服务时，要多注意对方的家庭结构是否存在特殊性。提示志愿者在写就书信时对对方的家庭结构有充分的了解。

建议 3：在提供课外书籍或者杂志寄送志愿服务时，要多考虑对方的抚养情况，避免对对方的情感造成刺激。

建议 4：在提供学习方面的经验或者建议时，无需过多考虑对方家庭劳动力和主要收入方式的因素，应该多考虑对方的家庭结构，从这点出发给出针对性的建议。

希望以上建议能够帮助“一帮一”活动更好地开展、延续、发展下去。

6 研究不足与展望

虽然我们得到了一些有用的结论，但是本研究还是存在相当的不足之处。

（1）研究所用的数据指标比较缺乏，缺乏如月薪、年收入等经济方面的定量指标，只有通过患病情况和收入来源情况推测出的劳动力情况；

（2）用李特克五点法评价的学习成绩是由孩子自己填写的，存在一些心理学上的认知误差；

（3）缺乏孩子本身对当前各指标的评价。

针对上述不足，我们对本项研究提出一些展望。

一是希望能够完善问卷。考虑到孩子年龄尚小，对家庭经济情况可能没有概念或者无法做出准确的估计，这一部分或许可以从多模态的配套问题中获得答案（比如线下家访记录）。

二是希望能够通过与当地有关部门合作，获得量化的经济信息和学习成绩等，这样获得的数据更加准确和真实。

参考文献

[1] 中国统计年鉴 2021，国家统计局, http://www.stats.gov.cn/tjsj/ndsj/2021/indexch.htm

[2] 中国儿童福利与保护政策报告 2019，北京师范大学中国公益研究院

[3] 家庭结构缺失对子女教育获得的影响，龙莹等, 2020

[4] 歧视知觉对留守青少年自伤的影响：愤怒的中介作用及性别差异，丁倩等, 2022

[5] 不同致贫原因脱贫人口的患病特征比较，李惠文等, 2021

Abstract

The poverty alleviation work of China has been completed successfully. How to continue assisting students in poverty-stricken areas in a targeted manner, especially how to make the effect of educational poverty alleviation stable and sustainable is an important topic in post-poverty era. This research uses two statistical analysis tools, SPSS and Python, to analyze the questionnaire results of students in two original poor areas and analyze the relationship among the metrics of these students, and provide some advice toward the educational poverty alleviation work after and “one helps one” work.