基于人寿保险公司数据的 SPSS 软件单因素方差分析

胡晨曦 杨杉 李蕊
四川大学锦城学院 计算机与软件学院 四川 成都 611731

【摘 要】本文针对不同年龄阶段与保费是否有显著性差异进行特征分析，运用 SPSS 分析软件中的单因素方差分析的方法进行验证，结合统计得出的数据进行分析后，证明两者间存在显著性差异，且年龄在 18-34 岁的保费是偏高于其余年龄段的，年龄在 60 岁以上的保费是偏低于其余年龄段，针对这一现象提出了合理的解释与建议。

【关键词】保险；SPSS 软件；单因素方差分析

1 引言

现如今中国老龄化、大病重疾、天灾人祸轮番上演的情况越来越多，我们的政府逐渐意识到单靠社保无法实现老有所养，单靠医保无法保证病有所医，单靠工伤保险无法保证意外之后有所依靠。只有依靠商业保险，才能用极微小的代价换来最及时有效的回报。随着中国人民生活发展水平的不断提高，保险已经走进了千家万户，伴随中国保险发展的几十年，人们的保险意识正不断提高，人们已经开始意识到了保险的重要性，并且主动去了解保险、选择符合自身情况的保险产品。特别是人们对于自身健康以及财产重视程度的不断增加，保险行业已经进入了高速发展的快车时代 [1]。本文利用SPSS软件对一份包含90万条数据的新投保数据做分析。结合分析出的结果，可以为保险公司和客户提供准确的参考数据，保险公司为了寻求更长远的发展，精准对客户进行定位分析，所以要分析市场需求，有针对性的提供更符合客户期望的保险产品。而客户也能通过分析得出的结果，结合自身的实际情况，选择最适合自己的保险产品。所以对于新投保数据进行分析有极大的价值。

2 数据准备及清洗

2.1 数据来源说明

数据来源于四川人寿保险公司的新投保数据。分别记录了机构、险种、投保时间、缴费、缴费期限、投保份数、总保费、保额、客户号、性别、年龄、婚姻、过去三年平均年收入、教育程度、职业、家庭人口共计 16 个字段，总共 90 万条数据，因为数据量特别的庞大，所以需要用到专门的数据分析软件 SPSS 来进行分析。

2.2 数据清洗

由于新投保的 90 万条数据集包含大量的数据，可能会有一定的脏数据会对我们的分析研究结果产生一定的影响。所以我们需要做数据清洗。数据清洗的目的在于删除重复值、纠正现有的错误，提供一致的结构化数据，以保证数据的正确和整洁。因而我们通过运用 Excel 工具的定位条件判断是否有空值并进行删除。通过对年龄进行升序排序，运用高级筛选功能对每一列的数据进行查看，查找出不合理的数据值做删除处理，例如年龄为 1 岁、婴幼儿、年收入却为 30000 元，对于教育程度都是无，家庭人口都是 0 的情况，我们将这两列没有意义的数据删除。通过简单的数据清洗后可以从数据中提取出更有价值的信息。

3 不同年龄保额特征分析

3.1 探究分析的前提条件

在具体的分析试验中，我们针对要考察的指标将其称为试验数据指标，对于影响试验数据指标的某个条件称为因素，把因素所处的状态称为发展水平，若在分析试验中只有这样一个重要因素改变则称为单因素分析试验。单因素方差分析被用来检验多重平均数之间是否存在差异，通过对数据变异的分析来推断两个或多个样本均数所代表的总体均数是否有差别的一种统计推断方法。通俗易懂的来讲，就是用来检验同一个影响因素的不同水平对变量是否有影响的一种方法。因为单因素进行方差分析是判断结果是否有显著性影响的一种发展统计数据分析的方法，即这种方法对于研究不同年龄段对于保额是否有显著性差异具有良好的判断效果 [2]。具体前提是每个样本必须是独立的随机样本，即样本之间不会相互影响，每个样本来自正态分布总体；而总体方差相等，即方差是齐次的。本文探索的是不同年龄阶段对于保费是否有显著性差异，由于年龄阶段都是独立的样本，且不会相互影响，每个样本都来自正态分布总体，方差相等，综上满足上述条件，所以本文将采用单因素方差分析的方法对此问题进行详细的分析。

3.2 分析操作过程

本文研究不同年龄与保额是否有显著性差异：首先对年龄、保额进行描述统计，发现年龄跨度比较大，如果对每一个年龄均进行分析，那么数据量太大，也不能

准确的得出想要研究的结果。因此需要对年龄进行分段处理。分段依据为：18 岁以下为一组，18–34 岁为一组，35–59 岁为一组，60 岁以上为一组。

基于分析目录下比较平均值选择单因素方差分析：将保额设置为因变量列表，把年龄设置为因子。并且在事后比较中的假定等方差一栏勾选 LSD，在不假定等方差一栏勾选塔姆黑尼 T2，在原假设检验一栏勾选使用与选项中的设置相同的显著性水平。在选项中的统计一栏勾选方差齐次性检验，在缺失值一栏勾选按具体分析排除个案，并将置信区间设置为 0.95，在对比栏中将其系数分别设置为 1 –1 –1 –1。

3.3 统计结果分析

通过上述操作后我们得到分析后的方差齐次性检验的结果如下图 1：

方差齐性检验

		莱文统计	自由度 1	自由度 2	显著性
保额	基于平均值	1364.024	3	900643	.000
	基于中位数	628.939	3	900643	.000
	基于中位数并具有调整后自由度	628.939	3	892744.095	.000
	基于剪除后平均值	1257.978	3	900643	.000

图 1（方差齐性检验结果）

本文采用“方差齐性检验”对检测样本中年龄与保额是否有显著性差异进行验证。该检验的原假设是年龄与保额有显著性差异。当所在总体的方差相等，在显著性差异水平取 0.05 时，若得到的显著性分析结果大于显著性影响水平，则接受原假设，反之则拒绝原假设 [3]。一般而言，我们进行方差分析的数据均应当满足方差齐性假设，即年龄与保额之间存在差异显著性的值，且该值要大于 0.05 方可进行 LSD 的分析，但由于我们的数据量有 90 多万条，结合图 1 的结果我们发现：本次齐性检验的结果为 0.000，小于 0.05（置信度为 95%），方差不具有齐次性，所以不能参考 LSD 的结果，应当以采用塔姆黑尼 T2 的结果为准。

综上结合分析出来的结果，我们发现方差不具有齐次性，所以我们要拒绝原假设，在事后检验多重性比较的图中我们要选择塔姆黑尼 T2 作为我们的判断参考依据，通过分析得出塔姆黑尼 T2 的数据分析结果如下图 2:

多重比较

因变量：保额

	(I) 年龄组	(J) 年龄组	平均值差值 (I-J)	标准误差	显著性	95% 置信区间 下限	上限
塔姆黑尼	1	2	-3800.38169*	685.67687	<.001	-5607.7194	-1993.0439
		3	-153.49563	684.29426	1.000	-1957.2157	1650.2245
		4	776.25872	695.31336	.842	-1056.2977	2608.8151
	2	1	3800.38169*	685.67687	<.001	1993.0439	5607.7194
		3	3646.88606*	61.35376	.000	3485.4612	3808.3109
		4	4576.64041*	137.71822	.000	4214.2860	4938.9948
	3	1	153.49563	684.29426	1.000	-1650.2245	1957.2157
		2	-3646.88606*	61.35376	.000	-3808.3109	-3485.4612
		4	929.75434*	130.66055	<.001	585.9668	1273.5419
	4	1	-776.25872	695.31336	.842	-2608.8151	1056.2977
		2	-4576.64041*	137.71822	.000	-4938.9948	-4214.2860
		3	-929.75434*	130.66055	<.001	-1273.5419	-585.9668

*. 平均值差值的显著性水平为 0.05。

图 2（多重比较结果）

我们对以上图 2 的结果进行分析发现：在组 1 中，在年龄组 2 的显著性值小于 0.05，说明有显著性差异，以组 1 和组 2 的平均值差值为负数举例，说明年龄在 18 岁以下的保额水平低于年龄在 18–34 岁之间的保额水平。在组 2 中，组 1，3，4 的显著性值都小于 0.05，说明有显著性差异，且组 2 与其他几个组的平均值差值的结果最后都为正数，说明组 2 的保额水高于年龄组 1，3，4。即年龄在 18–34 岁的保额水平高于其他年龄阶段的保额水平。在年龄组 3 中，组 2、4 的显著性值均小于 0.05，有显著性差异。通过查看平均差值得到，组 3 大于组 4、小于组 2。即说明了年龄组在 35–59 岁的保额水平高于年龄 60 岁以上的，小于 18–34 岁的保额水平。同理可得，在年龄组 4 中，组 2、3 的显著性值都小于 0.05，说明有显著性差异。说明年龄高于 60 岁以上的保额水平低于年龄段在 18–34 岁、35–59 岁的。综上所述，根据划分的年龄段来看，年龄在 18–34 岁的保费是偏高于其余年龄段的，年龄在 60 岁以上的保费是偏低于其余年龄段的。

4 结论

通过研究使用单因素进行方差分析的方法来探究：年龄与保额之间是否存在显著性差异。本文总结出了相应的结论以及合理的建议。统计数据分析在各学科中都扮演了一个重要的角色， 如果要进行正确的数据分析，在分析前必须要构建出正确的统计分析思想 [4]。方差分析是对检验数据进行分析，检验方差相等的多个正态总体的均值是否相等，进一步判断各因素对试验指标的影响是否明显，运用单因素方差分析的方法，可以区分出在不同的试验条件下，是否是影响试验结果的显著性因素，通过进一步的两两比较可以区分出不同实验条件间彼此的差异情况。 通过统计分析得出结论：不同年龄段的保额水平有显著性差异，年龄在 18–34 岁的保费是偏高于其余年龄段的，年龄在 60 岁以上的保费是偏低于其余年龄段的。

【参考文献】

[1] 周铭 . 基于灰色关联度分析法的企业价值评估——以我国保险企业为例 [J]. 广西质量监督导报 ,2020(08):112–113.

[2] 刘浩 , 史雨梅 , 余晓美 . 基于 SPSS 单因素方差分析在专业认同研究中的应用 [J]. 经济研究导刊 ,2020(05):71–73.

[3] 牛凯 . 数据分析之单因素方差分析 [J]. 产业与科技论坛 ,2019,18(02):57–58.

[4] 何成 .SPSS 中单因素方差分析的运用 [J]. 农业网络信息 ,2018(01):92–94.

[5] 范琦 . 消费者涉入、感知价值对商业健康保险购买意愿的影响研究 [D]. 成都：西南财经大学 ,2019.

[6] 王丽蕊 . 浅谈我国人寿保险营销存在的问题及策略完善 [J]. 商业经济 ,2020(12):49–50+93.