专利数据 SPSS 分析报告

信息分析课程期末项目

吴咏真

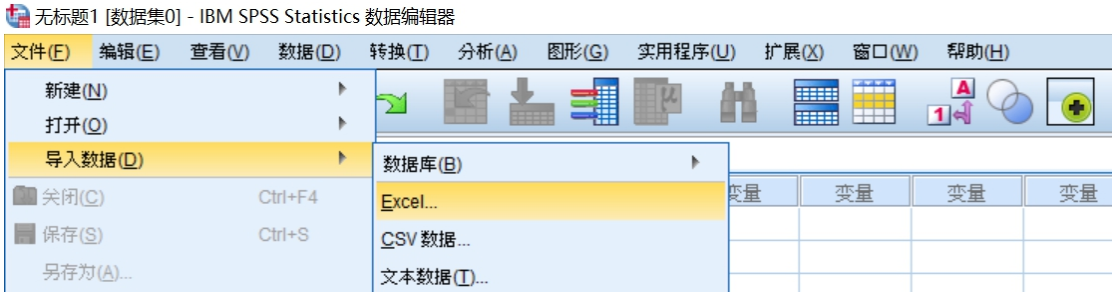
2020 年 7 月 9 日

目录

数据初探

导入 SPSS

在老师给定的表中完成数据录入后，我新建工作表，将自己的部分复制粘贴搬运过来，不含变量说明一共是 320 行，并删除姓名列。简单检查后启动 SPSS，导入数据，操作如下图所示。



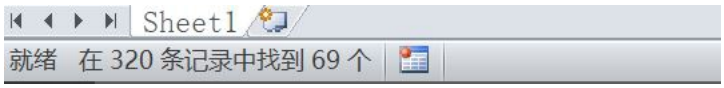
如果我希望能够对初始数据的构成（即频数统计/占比统计）情况有一定了解，我们可以：

- 在 SPSS 中通过分析→描述统计→频率实现频率表功能；

A类引证数量

		频率	百分比	有效百分比	累积百分比
有效	0	251	78.4	78.4	78.4
	1	12	3.8	3.8	82.2
	2	14	4.4	4.4	86.6
	3	12	3.8	3.8	90.3
	4	9	2.8	2.8	93.1
	5	8	2.5	2.5	95.6
	6	10	3.1	3.1	98.8
	7	3	.9	.9	99.7
	9	1	.3	.3	100.0
	总计	320	100.0	100.0	

- 在 EXCEL 中通过选中列→排序和筛选→筛选→选定数值（可多选），EXCEL 将显示符合要求的数据，并对结果进行计数。



问题说明

对于引证 通过上述的频数操作我们不难发现只有 A 类引证的样本量相对充

裕（69），X 类（7）和 Y 类（4）引证样本量小且取值单一，不满足研究条件。故对于第二问来说，我只能回答 A 类引证次数与寿命的相关性这一小问。

公开号	寿命	A类引证数	X类引证
CN1025220	2.24	0	3
CN1020413	3.18	2	1
CN1040173	5.28	0	1
CN1023932	2.52	5	1
CN1025062	2.39	4	1
CN1022604	2.80	1	2
CN1025219	2.20	4	1

公开号	寿命	A类引证数	X类引证数	Y类引证
CN1023271	2.47	1	0	3
CN1024192	2.52	3	0	2
CN1023209	2.66	0	0	1
CN1033029	5.60	2	0	2

对于被引 因为我的 320 条数据中只有 3 条数据有被引（如下图所示），样本量过小且取值单一，因此第三问即被引次数与寿命的相关性部分无法操作。第一问中包括的第二个小问题即“没有发生”被引的与“发生了”被引的寿命长度的差异也因为数据本身的原因而不能作答。

公开号	寿命	A类引证数	X类引证数	Y类引证数	专利权终	A类被引	首次A类被	X类被引数	首次X类被	Y类被引数	首次Y类被引年份
CN1023271	2.47	1	0	3	2014	1	2012	1	2012	0	
CN1020413	3.18	2	1	0	2014	1	2014	1	2013	0	
CN1031965	2.00	0	0	0	2015	1	2013	0		1	2013

综上所述，本报告将回答的问题是：

- 第一部分中：“没有发生”引证的与“发生了”引证的寿命长度的差异；
- 第二部分中：A 类引证次数与寿命的相关性。

第一问

试通过合适的差异分析探讨：“没有发生”引证的与“发生了”引证的寿命长度的差异。

单一样本 T 检验

单样本 T 检验用来分析样本均值与总体均值的差异，以此来判断这个样本来自总体的均值是否等于（大于或小于）某个已知总体的均值，适用条件是样本呈正态分布。典型的案例如：假定大学生的平均体重为 50 千克，现在某高校随机抽取 590 名大学生并检测起体重数据，问该校大学生的体重与一般大学生是否存在差异？对于我们的问题而言，不妨先研究“发生了”引证的寿命长度与总体数据的相比是否存在差异？

数据准备 我们将这 320 条数据的引证情况统计如下，把发生了引证的另存备用。此时只考虑 0 和 1 的区别。在总计的计数上，可以采取函数一步到位，笔者则采用了一个比较傻瓜的方法，因为这三类引证是“或”的关系（即只要其中

一类出现了引证，就算该专利发生引证），我们先在工作表中新建一列取名为“是否发生引证”，将 A 类引证一列反选去零，把被筛出的 69 行的“是否发生引证”单元格都填为“是”（可用自动填充功能），接着对 X 类、Y 类如法炮制，查漏补缺，最后对新建的这一列进行筛选，对“是”计数便可得出结果。

是否发生引证 / 类别	是	否
A 类引证	69	251
X 类引证	7	313
Y 类引证	4	316
总计	72	248

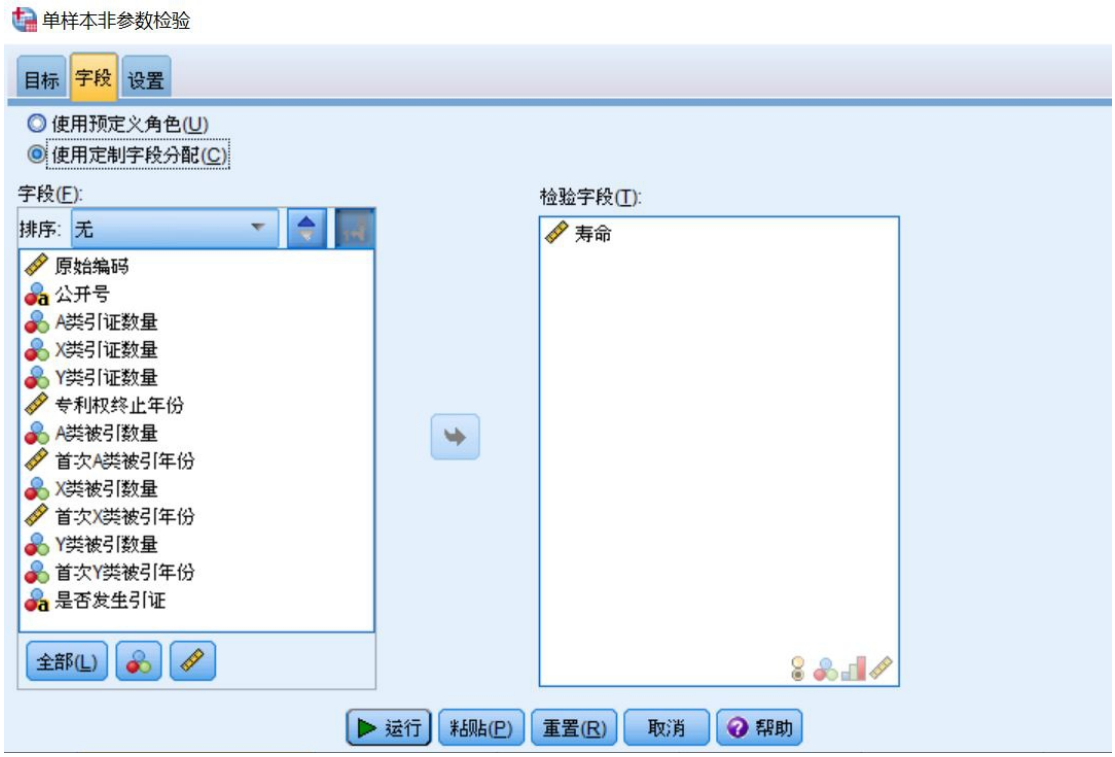
计算总体均值 打开 SPSS，针对“寿命”变量，在上方工具栏中单击“分析”、“描述统计”、“描述”，在选项中勾选平均值，得出总体均值。可用 EXCEL 验证演算结果是否一致。

描述统计

	N	均值
寿命	320	3.656115998
有效个案数（成列）	320	

平均值: 3.656115998　计数: 321　求和: 1169.957119

样本正态性检验 我们将发生引证的记录表导入软件中，完成样本数据的准备工作。选择“分析”、“非参数检验”、“单样本”，将寿命选为检验字段，单击运行。



假设检验摘要

	原假设	检验	显著性	决策
1	寿命 的分布为正态分布，平均值为 3.299138050341060，标准差为 .693384760126550。	单样本柯尔莫戈洛夫-斯米诺夫检验	.030[a]	拒绝原假设。

显示了渐进显著性。显著性水平为 .050。

a. 里利氏修正后

我们发现样本数据不符合正态分布。但是不必太过担心，单样本 T 检验法比较稳健，只要不是严重偏态都可以使用。此时我们需要通过描述法即描述数据偏度（K）和峰度(W)系数来检验数据的正态性。

系数	说明
偏度（skewness）	是统计数据分布偏斜方向和程度的度量，是统计数据分布非对称程度的数字特征； 正态分布的偏度为 0，两侧尾部长度对称。若以 bs 表示偏度，bs＜0 称分布具有负偏离，也称左偏态；bs＞0 称分布具有正偏离，也称右偏态；而 bs 接近 0 则可认为分布是对称的。
峰度（Kurtosis）	与偏度类似，是描述样本中所有取值分布形态陡缓程度的统计量； 这个统计量需要与正态分布相比较，峰度为 0 表示该数据分布与正态分布的陡缓程度相同；峰度大于 0 表示该数据分布与正态分布相比较为陡峭；峰度小于 0 表示该数据分布与正态分布相比较为平坦，峰度的绝对值数值越大表示其分布形态的陡缓程度与正态分布的差异程度

	越大。

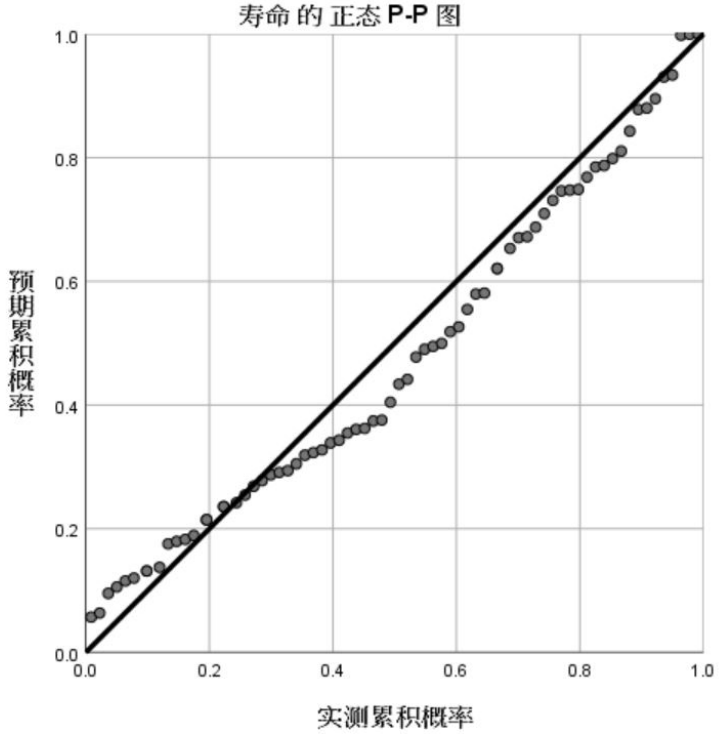
理论上来说，标准正态分布偏度和峰度均为 0，但现实中数据往往无法满足标准正态分布，因而如果峰度绝对值小于 10 并且偏度绝对值小于 3，则说明数据虽然不是绝对正态，但基本可接受为正态分布。

具体步骤是单击“分析”、“描述统计”、“描述”，选择需要检验的变量为寿命，在选项窗口中勾选峰度和偏度。从输出的结果上看，寿命的峰度绝对值为 2.371＜10，偏度的绝对值为 1.261＜3，所以基本可接受数据为正态分布。

描述统计

	N	最小值	最大值	均值	标准 偏差	偏度		峰度	
	统计	统计	统计	统计	统计	统计	标准 错误	统计	标准 错误
寿命	72	2.202739726	5.600000000	3.299138050	.6933847601	1.261	.283	2.371	.559
有效个案数（成列）	72								

我们可以用 P-P 图（根据变量的累积概率对应于所指定的理论分布累积概率绘制的散点图）较为直观地看一下，点击“分析”、“描述统计”、“P-P 图”，选择寿命为需要检验的变量并把检验分布选择成正态分布。不难看出，数据散点基本落在原点出发的 45°线附近，亦可知确实稍有偏态，但并不严重。



单一样本 T 检验 准备就绪，我们选择“分析”、“比较平均值”、“单样本 T 检验”，在检验值框中输入之前得到的总体均值，其他保持不变，点击确定。样本中包含 72 个个案，平均值约为 3.3，而总体均值约为 3.7。由下表可知显著性 P=0.000＜0.05，因此我们的结论是“发生了”引证的专利寿命长度与总体专利寿命长度存在显著差异。这符合我们对数据的主观判断和先前预期。

单样本统计

	个案数	平均值	标准 偏差	标准 误差平均值
寿命	72	3.299138050	.6933847601	.0817161776

单样本检验

检验值 = 3.656115998

	t	自由度	Sig.（双尾）	平均值差值	差值 95% 置信区间	
					下限	上限
寿命	-4.369	71	.000	-.356977948	-.519915378	-.194040517

独立样本 T 检验

在这一部分我们真正去研究“没有发生”引证的与“发生了”引证的寿命长度是否存在差异。

样本正态性检验 同上，我们需先验证“没有发生”引证的专利数据是否符合正态分布，操作过程在此处不再赘述，得出的结论是：“没有发生”引证的专利数据也不符合正态分布，但样本量相对大一些且偏态不严重，总体而言不影响后续分析。

假设检验摘要

	原假设	检验	显著性	决策
1	寿命 的分布为正态分布，平均值为 3.759754757287110，标准差为 1.095626313063150。	单样本柯尔莫戈洛夫-斯米诺夫检验	.000[a]	拒绝原假设。

显示了渐进显著性。显著性水平为 .050。

a. 里利氏修正后

统计

寿命

个案数	有效	248
	缺失	0
偏度		1.782
偏度标准误差		.155
峰度		5.622
峰度标准误差		.308

方差齐性检验与独立样本 T 检验 我们需先配置数据，将新建的“是否发生引证”列作为后面要用的分组变量，发生了引证的记为 1（把“是”改为数字 1），没有发生引证的记为 2（把空白处填上数字 2）。导入修改好了的 320 条记录，点

击“分析”、“比较平均值”、“独立样本 T 检验”，把寿命选为检验变量，是否发生引证选为分组变量，并定义组，如下图所示：

结果分析 查看结果，应当明确传统的独立样本 T 检验要求两样本方差齐，我们注意输出的“莱文方差等同性”一栏，发现显著性为 0.011＜0.05，说明**方差不齐**，此时我们的 T、P 值应取下面一行，即 t=-4.292，P=0.000。又因为 sig 值＜0.05，我们得出的结论是：“没有发生”引证的与“发生了”引证的寿命长度有显著差异。这符合我们的猜想，“前有古人”的专利在创新性上往往不及“开天辟地”的专利，由统计数据亦可看出“发生了”引证的寿命均值（约 3.3）明显短于“没有发生”引证的寿命均值（约 3.8）。

[数据集4]

组统计

	是否发生引证	个案数	平均值	标准 偏差	标准 误差平均值
寿命	1	72	3.299138050	.6933847601	.0817161776
	2	248	3.759754757	1.095626313	.0695723405

独立样本检验

		莱文方差等同性检验		平均值等同性 t 检验						
									差值 95% 置信区间	
		F	显著性	t	自由度	Sig.（双尾）	平均值差值	标准误差差值	下限	上限
寿命	假定等方差	6.477	.011	-3.374	318	.001	-.460616707	.1365032035	-.729180201	-.192053213
	不假定等方差			-4.292	183.519	.000	-.460616707	.1073212199	-.672358769	-.248874645

注：我们也可以采取“分析”、“描述统计”、“探索”，将寿命选入因变量列表框，将是否发生引证选入因子列表框，输出勾选“两者都”，在“图”中勾选“因子级别并置”和“未转换”来得到方差齐性检验结果；或者采用比较均值下的单因素 ANOVA，选项窗口选择方差同质性检验。总之，条条大路通罗马，得到的运算结果也是一样的（0.011）。

方差齐性检验

		莱文统计	自由度 1	自由度 2	显著性
寿命	基于平均值	6.477	1	318	.011
	基于中位数	5.906	1	318	.016
	基于中位数并具有调整后自由度	5.906	1	289.346	.016
	基于剪除后平均值	6.084	1	318	.014

卡方检验

卡方检验针对分类变量，因此我们需要重新整理数据。

整理数据 不妨用简单粗暴的方法将所有记录分为两组：高寿命（寿命大于或等于总体均值）和低寿命（寿命小于总体均值）。在 EXCEL 中通过 Ctrl+Shift+L 组合键开启对各个列的筛选功能。对寿命列利用数字筛选自定义大于或等于 3.656115998/小于 3.656115998；再分别将是否发生引证列选为 1/2，填写完成后得到下表：

组别	高寿命	低寿命	合计
发生引证组	19	53	72
未发生引证组	119	129	248
合计	138	182	320

可以说，如果寿命的长短与是否发生引证没有关系，那么以蓝底色单元格为例，它的数值应该是 138*72/320=31.05（期望值），但实际值为 19，说明理论和实际之间有差距。卡方检验就是统计样本的实际观测值与理论推断值之间的偏离程度的，实际观测值与理论推断值之间的偏离程度就决定卡方值的大小。卡方值越大，二者偏差程度越大；反之，二者偏差越小；两个值若完全相等，卡方值就为 0。

录入数据 接下来我们将四格表中的数据录入 SPSS，并在变量视图设置好相应的值标签：引证组别中 1 代表发生引证、2 代表未发生引证；寿命长短中 0 代表低寿命、1 代表高寿命。在设置完成以后我们也可以通过点击工具栏上面的值标签按键查看我们的值标签是否有误。

	记录数目	引证组别	寿命长短	变量	变量
1	19	1	1		
2	53	1	0		
3	119	2	1		
4	129	2	0		
5					

	名称	类型	宽度	小数位数	标签	值	缺失	列	对齐	测量	角色
1	记录数目	数字	15	0		无	无	8	右	标度	输入
2	引证组别	数字	15	0		{1, 发生引证...	无	10	右	标度	输入
3	寿命长短	数字	15	0		{0, 低寿命}...	无	9	右	标度	输入

频数加权 点击“数据”、“个案加权”，选择对记录数目进行加权，做这一步主要是因为不进行加权而直接进行卡方分析是很容易出错的，将影响到整个分析结果。

卡方检验 单击工具栏中的“分析”、“描述统计”、“交叉表”，将行选为引证组别，列选为寿命长短，在右边的统计选项中勾选卡方，其他地方保持系统默认，确定后输出结果。

结果分析 一共输出了三张表，其中前两张主要是对数据样本量进行了统计和整理，我们重点看第三张。由于这个案例中所有的理论数 T≥5 并且总样本量 n≥40，故用 Pearson 卡方进行检验。如图，Pearson 卡方值为 10.609，对应的显著性 P 值为 0.001＜0.05，说明本次实践中两个组别（发生引证、未发生引证）的寿命长短构成是有差别的。卡方检验其实是我一时兴起的尝试，但某种意义上也算是回答了这一问的问题。

卡方检验

	值	自由度	渐进显著性（双侧）	精确显著性（双侧）	精确显著性（单侧）
皮尔逊卡方	10.609[a]	1	.001		
连续性修正[b]	9.747	1	.002		
似然比	11.047	1	.001		
费希尔精确检验				.001	.001
线性关联	10.576	1	.001		
有效个案数	320				

a. 0 个单元格 (0.0%) 的期望计数小于 5。最小期望计数为 31.05。

b. 仅针对 2x2 表进行计算

第二问

试通过相关/回归分析论证引证次数与寿命的相关性。

双变量相关分析

为了进行两个变量之间的比较，通常可以使用相关分析的方法。Pearson 相关系数是一种线性关联度量，适用于两个变量的度量水平都是尺度数据，并且两个变量的总体是正态分布或接近于正态分布的情况。

数据准备 我们将发生了 A 类引证的 69 条数据单独建表并导入 SPSS，分别对 A 类引证数量和寿命这两个变量进行正态性检验(方法参见第一问，不再赘述)。发现它们都不是标准的正态分布，但偏态不严重，接近于正态分布，不影响处理。

统计

A类引证数量

个案数	有效	69
	缺失	0
偏度		.483
偏度标准误差		.289
峰度		-.562
峰度标准误差		.570

统计

寿命

个案数	有效	69
	缺失	0
偏度		1.274
偏度标准误差		.289
峰度		2.867
峰度标准误差		.570

Pearson 相关性检验 在“分析”菜单的“相关”子菜单中选择“双变量”，在弹出的对话框中将 A 类引证数量和寿命选入“变量”窗口，选择“皮尔逊相关系数”，然后确定。

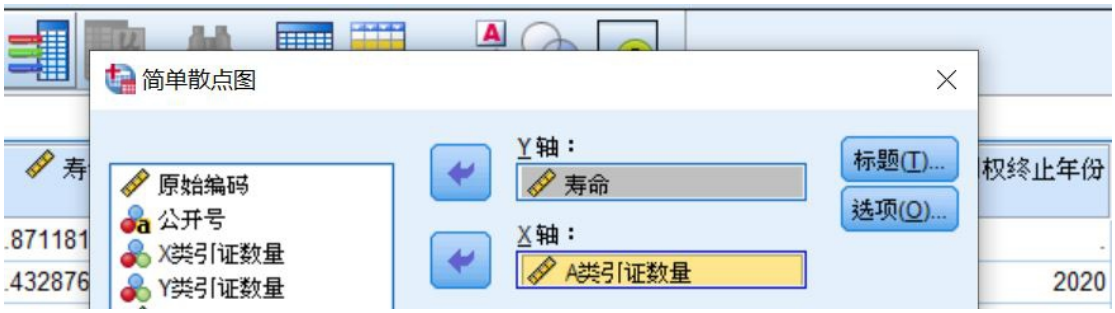
相关性

		A类引证数量	寿命
A类引证数量	皮尔逊相关性	1	-.255*
	Sig.（双尾）		.034
	个案数	69	69
寿命	皮尔逊相关性	-.255*	1
	Sig.（双尾）	.034	
	个案数	69	69

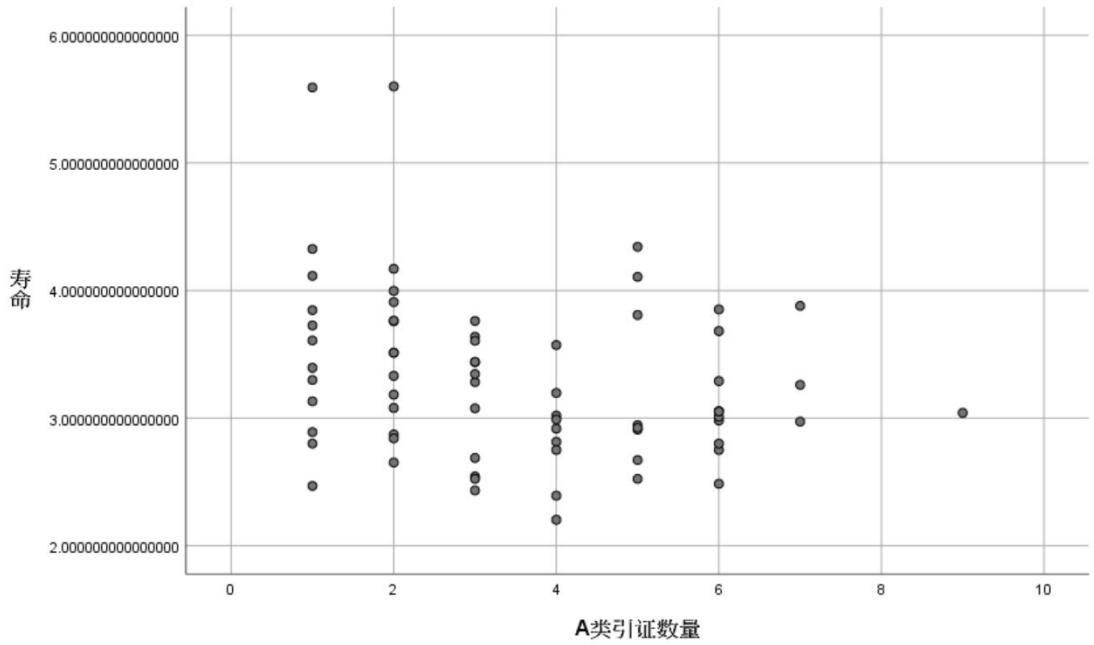
*. 在 0.05 级别（双尾），相关性显著。

结果分析 由输出的矩阵表可知，A 类引证数量与寿命的相关性系数为-0.255，在这个数据的旁边有一个星号，代表用户指定的显著性水平为 0.05 时，统计检验的相伴概率 0.01＜P＜0.05（在表格中的显示为 0.034）。结论是，专利的 A 类引证数量与其寿命的长短相关，且为负相关，即 A 类引证数量越多，寿命越短。（注：但相关性并不显著，可以被看作是**弱相关**，因为相关性系数的绝对值比较小。）

关于这一点，其实并不是不好理解，我们点击上方工具栏中的“图形”，在旧对话框中找到“散点图/点图”，弹出的对话框中我们把类型选为“简单散点图”，以寿命为 Y 轴、A 类引证数量为 X 轴画图。



我们会发现，二者之间其实**并不是线性关系**。而 Pearson 相关系数是用来反映两个变量线性相关程度的统计量，因此我们此前得出的结果不尽如人意也是可以理解的。



单因素方差分析

单因素方差分析（one-way ANOVA）也称为 F 检验，简单的来说，它就是用来检验同一个影响因素的不同水平对因量是否有影响的一种方法。

数据准备 我们在原表的基础上新增一列，命名为引证分组。将 A 类引证数量为 1 和 2 的（共计 24 条数据）记为低（赋值为数字 1）；3、4、5 的（共计 29 条数据）记为中（赋值为数字 2）；6、7、9 的（共计 14 条数据）记为高（赋值为数字 3）。

在这里我本来也尝试了使用 SPSS 分类下的 K-均值聚类功能，规定了生成 3 个簇，但结果不太好，高组的样本量太少（只有 4 条），导致后续分析上出现了一些问题，故而还是采取手动分类的版本。

每个聚类中的个案数目

聚类	1	27.000
	2	4.000
	3	38.000
有效		69.000
缺失		.000

引证分组

		频率	百分比	有效百分比	累积百分比
有效	低	26	37.7	37.7	37.7
	中	29	42.0	42.0	79.7
	高	14	20.3	20.3	100.0
	总计	69	100.0	100.0	

单因素方差分析 点击“分析”、“比较平均值”、“单因素 ANOVA 检验”，在弹出的选项卡中，将寿命选入到因变量列表中，将引证分组选入到因子中；点击右边的“事后比较”，在弹出的选项卡中选择 LSD 然后点击继续；再点击右边的“选项”，在弹出的选项卡中选择“描述”和“方差齐性检验”，其他保持默认不变，点击确定。

结果分析 在结果中，我们首先要看的就是方差齐性检验，因为如果不齐则不能使用单因素方差分析法，后面的结果是没有意义的。可见，显著性 P＞0.05，说明方差是齐的。

方差齐性检验

		莱文统计	自由度 1	自由度 2	显著性
寿命	基于平均值	1.954	2	66	.150
	基于中位数	2.009	2	66	.142
	基于中位数并具有调整后自由度	2.009	2	52.930	.144
	基于剪除后平均值	1.856	2	66	.164

在 ANOVA 表格中我们可以看到，P=0.011＜0.05，说明这三个组间至少有两个组之间是存在显著性差异的，即不同的 A 类引证次数对专利的寿命有影响。（注：这里我们并不能看出究竟有怎样的差异，还需要看下一个表格。）

ANOVA

寿命

	平方和	自由度	均方	F	显著性
组间	3.696	2	1.848	4.882	.011
组内	24.982	66	.379		
总计	28.678	68			

移步事后检验表格，我们可以看出中低两组间和高低两组间存在显著差异，而中高两组间无显著性差异（P=0.795＞0.05）。

事后检验

多重比较

因变量: 寿命

LSD

(I) 引证分组	(J) 引证分组	平均值差值 (I-J)	标准 错误	显著性	95% 置信区间	
					下限	上限
低	中	.492946300*	.1661635104	.004	.1611901473	.8247024524
	高	.440687765*	.2039479248	.034	.0334925950	.8478829348
中	低	-.492946300*	.1661635104	.004	-.824702452	-.161190147
	高	-.052258535	.2002217246	.795	-.452014106	.3474970362
高	低	-.440687765*	.2039479248	.034	-.847882935	-.033492595
	中	.0522585350	.2002217246	.795	-.347497036	.4520141061

*. 平均值差值的显著性水平为 0.05。

二元 Logistic 回归分析

使用 Logistic 模型前，我们需判断是否满足以下七个研究假设：

- 因变量即结局是二分类变量。
- 有至少 1 个自变量，自变量可以是连续变量，也可以是分类变量。
- 每条观测间相互独立。分类变量（包括因变量和自变量）的分类必须全面且每一个分类间互斥。
- 最小样本量要求为自变量数目的 15 倍，也有一些研究者认为样本量应达到自变量数目的 50 倍
- 连续的自变量与因变量的 logit 转换值之间存在线性关系。
- 自变量间不存在共线性。
- 没有明显的离群点、杠杆点和强影响点。

我们的打算是只设一个自变量（引证分组），而将因变量即寿命通过聚类功能分为两个水平，符合 Logistic 模型的适用条件。

数据准备 在“分析”、“分类”中选择“K-均值聚类分析”，将寿命拖进变量选框，制定聚类数为 2，点击右侧的“选项”，勾选要求显示每个个案的聚类信息，点击确认。

最终聚类中心

	聚类	
	1	2
寿命	3.986962908	2.925878410

最终聚类中心之间的距离

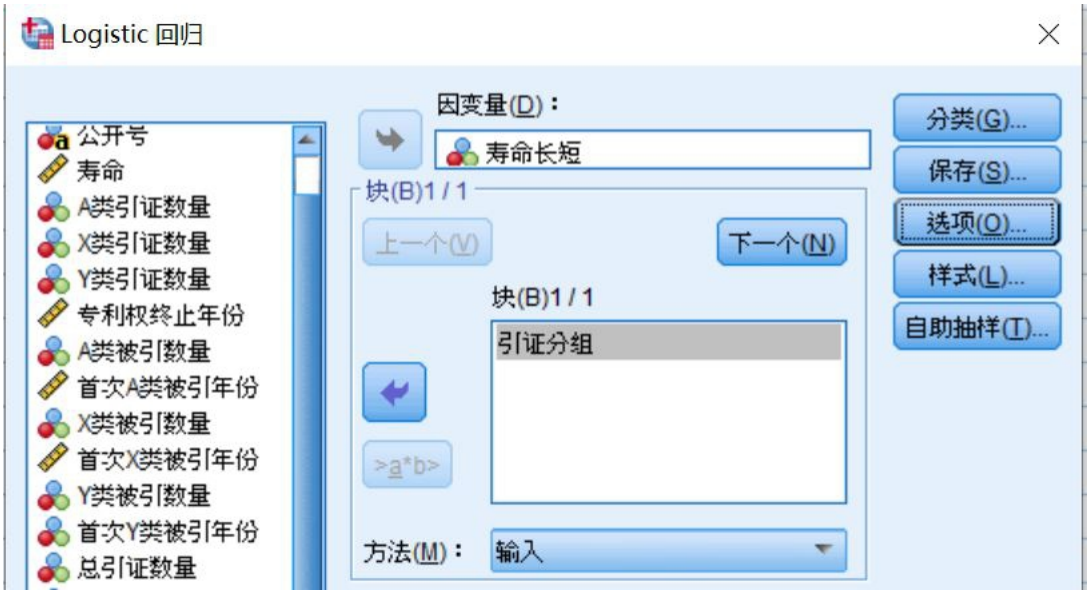
聚类	1	2
1		1.061
2	1.061	

每个聚类中的个案数目

聚类	1	24.000
	2	45.000
有效		69.000
缺失		.000

通过运行结果可知，组 1 代表高寿命、组 2 代表低寿命，每个类中的样本量也比较好看，初步判定可用。我们对照输出的聚类成员表新建一个变量，取名寿命长短，完成编码。

回归分析 在菜单栏中遵循“分析”、“回归”、“二元 Logistic”的路径打开配置框，把寿命长短选为因变量，自变量选择引证分组；在选项中勾选霍斯默-莱梅肖拟合优度。方法默认采用输入（将所有变量一次纳入到方程），运行得到结果。在输出的一长串表中我们挑选几个比较重要的进行解释。



结果分析 我们看到的第一个结果是对案例的描述，第一个列表告诉你有多少数据参与计算，有多少数据是缺失值；第二个列表告诉你因变量的编码方式，得分为 1 代表低寿命，得分为 0 代表高寿命。

个案处理摘要

未加权个案数[a]		个案数	百分比
选定的个案	包括在分析中的个案数	69	100.0
	缺失个案数	0	.0
	总计	69	100.0
未选定的个案		0	.0
总计		69	100.0

a. 如果权重为生效状态，请参阅分类表以了解个案总数。

因变量编码

原值	内部值
高寿命	0
低寿命	1

这个列表告诉我们在没有任何自变量进入以前，预测所有的记录都是低寿命（样本量大）的正确率，正确率为 65.2%。

分类表[a,b]

			预测		
			寿命长短		
	实测		高寿命	低寿命	正确百分比
步骤 0	寿命长短	高寿命	0	24	.0
		低寿命	0	45	100.0
	总体百分比				65.2

a. 常量包括在模型中。

b. 分界值为 .500

下面这个表格，通过显著性的值可以知道如果将模型外的变量纳入模型，则整个模型的拟合优度改变是否有统计学意义。由于 sig 值小于 0.05，说明有统计学意义。

未包括在方程中的变量

			得分	自由度	显著性
步骤 0	变量	引证分组	5.418	1	.020
	总体统计		5.418	1	.020

Omnibus 表是对模型系数的综合检验。其中模型一行输出了 Logistic 回归模型中所有参数是否均为 0 的似然比检验结果。P＜0.05 表示本次拟合的模型中，纳入变量的 OR 值有统计学意义，即模型总体有意义。

模型系数的 Omnibus 检验

		卡方	自由度	显著性
步骤 1	步骤	5.682	1	.017
	块	5.682	1	.017
	模型	5.682	1	.017

霍斯默-莱梅肖是用来检验模型的拟合优度。当 P 值不小于检验水准时（即 P＞0.05），认为当前数据中的信息已经被充分提取，模型拟合优度较高。

霍斯默-莱梅肖检验

步骤	卡方	自由度	显著性
1	1.002	1	.317

新的一张分类表，这里展示了使用该回归方程对案例数据进行分类，其准确度为 68.1%，相比之前提高了三个百分点，有进步但不明显。

分类表[a]

			预测		
			寿命长短		
	实测		高寿命	低寿命	正确百分比
步骤 1	寿命长短	高寿命	14	10	58.3
		低寿命	12	33	73.3
	总体百分比				68.1

a. 分界值为 .500

最后这张表显示，引证分组中赋值越高的研究对象（即 A 类引证次数越多），其寿命通常越低（系数为 0.867＞0，但因为在因变量赋值时比较变扭的把 1=低寿命，0=高寿命，所以其实是负相关）；相对于赋值较低的研究对象（A 类引证次数越少），赋值较高的研究对象（A 类引证次数越多）专利寿命低的风险是 2.379 倍。

方程中的变量

							EXP(B) 的 95% 置信区间	
		B	标准误差	瓦尔德	自由度	显著性	Exp(B)	下限	上限
步骤 1[a]	引证分组	.867	.384	5.099	1	.024	2.379	1.121	5.046
	常量	-.888	.695	1.636	1	.201	.411		

a. 在步骤 1 输入的变量：引证分组。

其他探索

针对现有的数据，我们还能做些什么？

Kaplan-Meier 生存分析

首先需要明确生存时间（Survival Time）的基本概念。它指的是指从某一起点到事件发生所经过的时间。生存是一个广义的概念，不仅仅指医学中的存活，也可以是机器出故障前的正常运行时间，或者下岗工人再就业前的待业时间等等。有的时候甚至不是通用意义上的时间，比如汽车在出故障前的行驶里程，也可以作为生存时间来考虑。

在我们的记录中，寿命一栏可以类比汽车的案例，它表明的是该专利自诞生以来到终止（如果尚未终止，则截止到目前即 2020 年）所经历的时间（单位：年）。因为专利颁布时间的不同，这些数据实际并不处在同一条起跑线上。

数据准备 我们在包含了全部 320 条记录的原表格后新增两列作为新变量，分别命名成是否发生引证和是否终止。前者中 1 代表发生，2 代表未发生；对于后者而言，专利权终止年份一栏为空值的记为 0，表示存续，该栏有具体年份数字的都记为 1，表示终止。可以通过“分析”、“描述统计”、“交叉表”看到每个类别下包含的样本数量。

是否发生引证 * 是否终止 交叉表

计数

		是否终止		
		存续	终止	总计
是否发生引证	发生	29	43	72
	未发生	145	103	248
总计		174	146	320

生存分析 点击“分析”、“生存分析”、“Kaplan-Meier”，在弹出的选框中将时间选为寿命、状态选为是否终止、因子选为是否发生引证，并点击定义事件，明确告诉软件工具本次数据研究的目标事件是数字 1，即指定“终止”是事件的结果。

接着打开比较因子选框，将检验统计的三个内容全部勾选。既然我们希望考察单个因素（是否发生引证的二分类变量）是否对生存产生影响，我们就需要比较这两组案例生存状态的差异。对这三种检验差异是否存在的方法简单介绍如下，笔者是 SPSS 小白，保险起见将它们全部选上。

秩的对数	对所有时点均赋予相同权重，该法对于生存分布的早期差别敏感。
布雷斯洛	用每个时点的历险数（暴露数）对时点加权，该法对于生存分布的后期差别敏感。
塔罗内-韦尔	用每个时点的历险数平方根对时点加权，当生存曲线或危险函数曲线有交叉时可选择此项。

点击“选项”，在图中勾选生存分析函数，它就是生存曲线图，最是让使用者关注的结果。配置完成后回到主页面点击确定，观察结果。

结果分析 如图，给出了各组生存时间均值和中位值，及其相应的标准误和 95%CI。此表可观测各组生存时间的集中趋势和离散程度。此处要强调，由于寿命数据不符合正态分布，所以平均值的相关估算结果并不是最重要的，主要以中位数（即中位生存时间）为准。在本案例中，我们可以看出，发生了引证的专利寿命中位数是 3.605 年，未发生引证的专利寿命中位数是 4.233 年，总体专利寿命中位数 3.978 年。

生存分析时间的平均值和中位数

是否发生引证	平均值[a] 估算	平均值[a] 标准 错误	平均值[a] 95% 置信区间 下限	平均值[a] 95% 置信区间 上限	中位数 估算	中位数 标准 错误	中位数 95% 置信区间 下限	中位数 95% 置信区间 上限
发生	3.693	.139	3.421	3.965	3.605	.171	3.270	3.941
未发生	5.337	.302	4.744	5.930	4.233	.337	3.572	4.894
总体	4.986	.250	4.496	5.475	3.978	.036	3.908	4.048

a. 如果已对生存分析时间进行检剔，那么估算将限于最大生存分析时间。

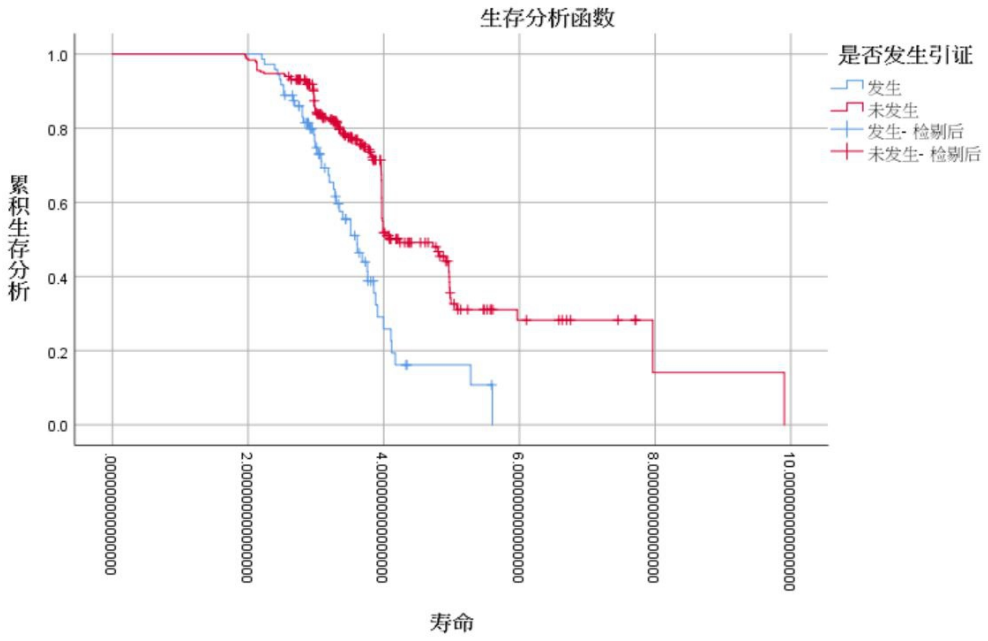
通过统计学检验给出了生存函数（指生存率）的组间比较。由假设检验的数据可知三种统计量结果基本一致，$p<0.05$，在本案例中我们便能认为发生引证将降低专利的“生存率”/寿命。

总体比较

	卡方	自由度	显著性
Log Rank (Mantel-Cox)	23.156	1	.000
Breslow (Generalized Wilcoxon)	16.647	1	.000
Tarone-Ware	20.194	1	.000

针对 是否发生引证 的不同级别进行的生存分析分布等同性检验。

生存分析函数图是以生存时间为横轴，生存率为纵轴绘制而成的阶梯形曲线，用于说明生存时间与生存率之间的关系。从图形上看，两条生存曲线在前期（寿命 2 年多的时候）出现了一次交叉，两组案例在前 4 年的累积生存率迅速下降（其中发生了引证的专利下降得更为陡峭），未发生组的曲线总体在上方，说明它们的寿命情况要比发生了引证的专利的乐观一些。这种差异，通过之前的总体比较，我们更倾向于是由是否发生引证这一事件造成的，而非误差。



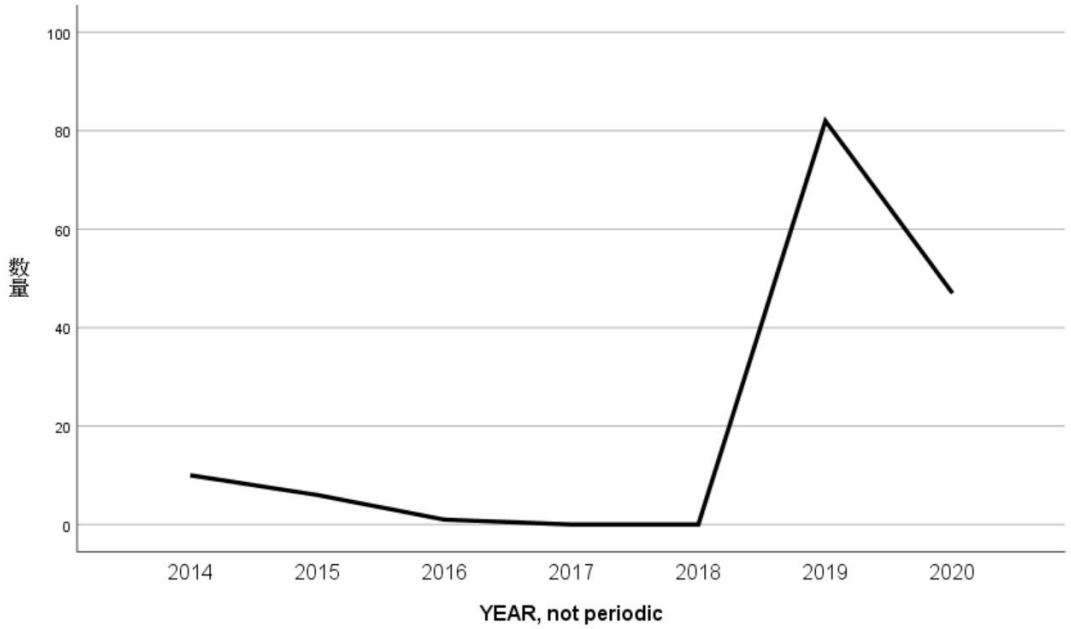
总结

这次作业是一次真正的、使用真实数据集的实操性作业，从获取、整理到预处理（这个步骤真的非常重要，你必须先认识你的数据，包括它的类型结构等，比方说 A 类引证次数，它虽然是数字型，但是我并不认为它可以直接被当做连续型变量）、分析数据并得出结论都是我一个人亲自完成的，体验感满分。

通过实践，我对 SPSS 的理解和运用相较之前也更为深入和熟练，同时认识到了一些操作上的误区（比如对方法适用的前提条件不够了解就直接硬着头皮上）；在得出有意义的结论时会发自内心感到快乐。

当然，美中不足是这次的数据总体不太好看，有价值的比较少（我认为遗憾的主要并不是有些既定问题无法回答，因为就像老师在任务说明里写的那样，大模块下的小问题多相通，既然能够研究 A 类引证次数与寿命的相关性，显然也能研究 X 类引证次数与寿命的相关性，在分析方法/操作步骤这样的核心环节上二者或许并无多少差异），无法研究不同类型引证/被引对寿命影响贡献度的差异性，我的部分“脑洞”亦无法实现。

譬如在专利权终止年份这一项上，我的数据是这样的（如下图），17 和 18 年甚至没有专利终止，而 19 年的终止数量暴增，我认为这是由于样本量小而导致的异常数据，如果记录数够多，或者时间被进一步明确到月/周/日，也许我们可以进行**时间序列预测分析**，看看是否具有周期性规律，大概估计一下未来专利“停摆”的情况。（注：笔者之前对全球疫情 COVID-19 数据进行分析时，就发现一周之中周日和周一确诊的人数最多。）



再比如，我们现在收集的数据，并未包括专利的具体内容信息，但这或许也

是影响其寿命的原因，一个假设是：医学类专利往往比日用类专利（比如，一种掏耳朵的工具）坚持的时间更久。

总之，我认为这会是一个很有意思的研究，如果样本量更大、对每个记录特征的提取更加丰富的话，应该能发现不少有趣的内容（当然，这都要建立在方法选择正确，参数配置合适的基础上），我很期待接下来的工作。