第二节 数据分析

完成数据清洗后，我们才能对数据进行分析。在市调中常使用到的数据列表形式有两种，一种是针对每个对象的单一变量列表，如受访者的年龄列表，这种列表有助于研究人员可以清晰地“看到”对象；这二种则是交叉列表。在交叉列表中包含有至少两个变量的数据或多个变量的数据，从而帮助研究人员对变量间的关系进行判断。

研究人员需要根据所要研究的问题选择合适的分析水平，根据所掌握的变量的不同类别（定类、定序、定距、定比）选择合适的统计分析方法（参考第四章测量）。本节我们将结合 SPSS 对一个变量、两个变量和多个变量进行分析，介绍相关的统计方法：

单变量——最简单的形式，通过单个变量来描述某种情况

双变量——子群比较，通过两个变量同时描述一个事件

多变量——同时分析三个或多个变量

9.2.1 单变量统计分析

用同一变量不同类别的属性构成对该变量的描述，是进行其他统计运算的基础环节。常见的例子如性别——男性的数量-女性的数量。研究人员在开展任何数据分析时，都应该首先对单变量的频数进行运算并查验数据录入是否正确。

对单变量的统计分析可分为两部分：频数分布，集中趋势测量。

频数

频数（Frequency）指某一个取值的个案数。通过计算一个变量下的不同取值的频数，汇总成频数表进行分析的方式一般适用于离散型变量。所谓离散型变量是指当变量的可能取值是一组自然数或整数时，变量为**离散型（Discrete variable）**，常见的定类变量和定序变量为离散型变量。对**连续型变量（Continuous variable，**可能取值无限连续，如定距变量）则必须先将变量的取值进行分组，每个分组作为一个新的选项，然后可以对这些新的选项进行频数表的分析和计算。

集中趋势测量与离散趋势测量

频数分布有两个特征：集中趋势和离散趋势。对其中一个特征的单独测量都很难代表整组数据的特征，因此集中趋势和离散趋势相结合才能更好地对数据的分布进行描述。

对数据集中趋势的测量，常用的统计量有：百分数（Presentation）、均值（Mean）、众数（Mode）、中位数（Median），表 9-2 是对相关统计方法的解释。

表 9-2 集中趋势测量

百分数	该取值的个案占总样本的比例。（频数/样本量）× 100%
均值	反映一组呈对称分布的变量值在数量上的平均水平。样本的所有 N 个观察值之和除以样本量。
众数	表示一组数据中出现次数最多或最常见的数值。 众数适合用于描述定类变量和定序变量。
中位数	表示一组数据按照大小的顺序排列时中间位置的那个数值，即针对某个变量，有 50%的个案的取值在中位数以下。适用于偏态分布资料和一端或两端无确切的数值的资料。

对数据离散趋势的测量，常用的统计量有：极差（Range）、四分位差（Quartile deviation）、方差（Variance）与标准差（Standard Deviation），表 9-3 是对相关统计方法的解释。

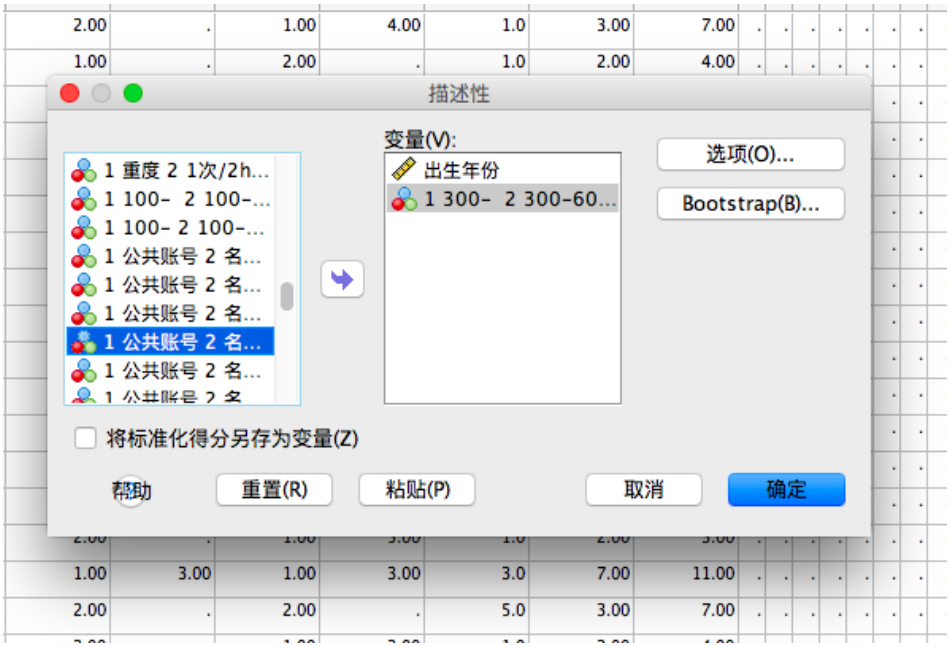
表 9-3 离散趋势测量

极差	最大值和最小值之间的距离，为一组数据的最大值和最小值之差，但极差不能反映所有数据的变异大小，且极易受样本含量的影响。常用以描述偏态分布。
四分位差	它是由第 3 四分位数与第 1 四分位数相减得到，常和中位数一起描述偏态分布，即 75%位置的取值减去 25%位置的取值所得。
方差与标准差	表示分布对平均数的偏离程度或伸展程度的度量。反映一组数据的平均离散水平，消除了样本含量的影响。

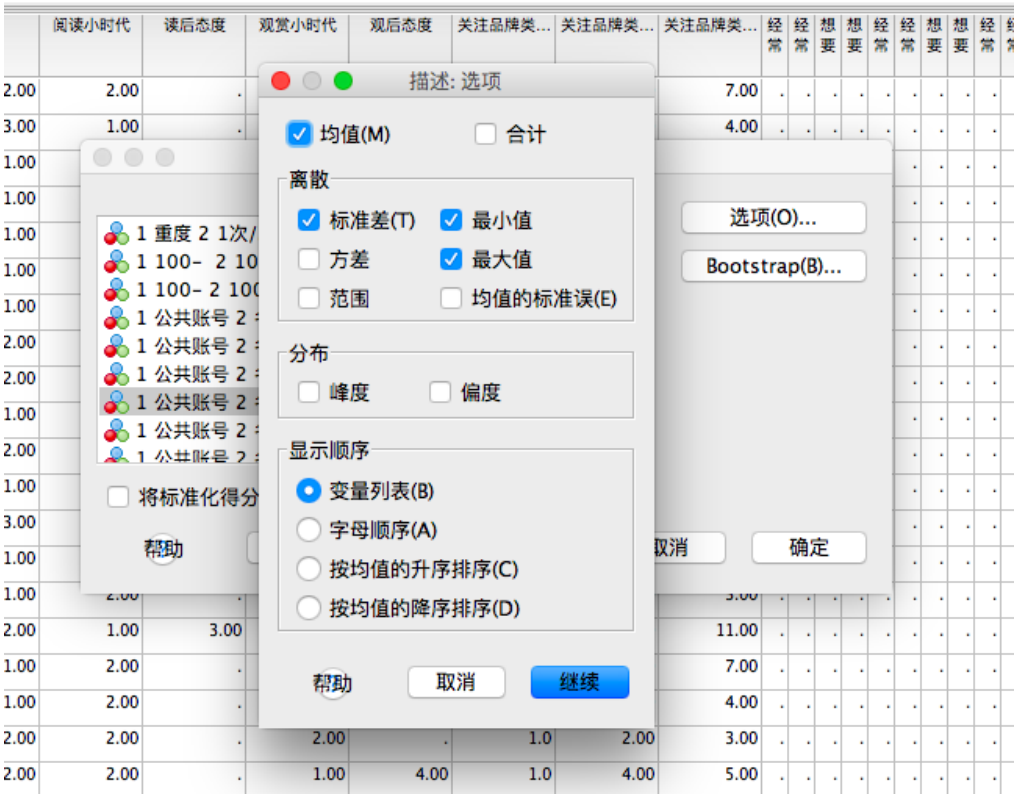
以上统计量均都可以通过公式计算获得，部分公式在表 9-2 和 9-3 中已经列举。接下来我们将通过 SPSS 统计软件直接获得它们的值：

SPSS 操作方法：从菜单中选择“分析”—“描述统计”—“描述”，弹出“描述性”对话框。

在所弹出的“描述性”对话框选中需要分析的变量后点击“→”加入分析框（**变量(V)**）。在将需要进行分析的变量从左边窗口选入右边变量窗口时，可一次选择多个变量，SPSS 将对这些变量进行逐一分析。



单击“描述性”对话框右侧“选项”按钮，在弹出的“描述：选项”对话框中勾选需要的统计量，并点击继续返回“描述性”对话框，单击确认。SPSS 会自动弹出输出窗口返回所需值。



细心的读者可能已经发现在描述统计菜单中并没有频数分析。在 SPSS 中频数分析在描述统计菜单下的“频数”中：“分析”→“描述统计”→“频率”。在弹出的频数分布对话框中将需要分析的变量从左侧源变量中选入右侧，并选择具体统计量后点击确定即可。具体步骤如下：

在分析菜单下，选择描述统计后选择“频率”

在弹出的频率对话框中，选择需要统计的变量

点选“统计量”、“图表”选项，在对应弹出框中选中所需要的统计量及图表，点击继续返回频率对话框，点击确定。

假设检验

在调研过程中，对数据集中趋势和离散趋势的度量是非常有效的。但是在实际操作中，研究人员往往还会根据对调查对象的了解和实际情况，提出一些想法，这些想法可以通过对所获得资料的统计分析加以求证，这一统计分析过程就是假设检验。**假设检验（Hypothesis testing）**又称显著性检验，是根据一定假设条件由样本推断总体的方法。假设检验的基本原理是小概率事件，即小概率事件在一次观察中不应该出现，如果出现了，则可以拒绝原假设。假设检验的逻辑如下：

第一，问题是什么？根据所需要研究的问题，提出原假设（记作 H_0）和备择假设（记作 H_1）。

问题：每周三看电影的人数与比其他时间看电影的人数比较

原假设 H_0：周三看电影的人数不比其他时间看电影的人少

备择假设 H_1：周三看电影的人数比其他时间看电影的人少

原假设和备择假设在逻辑上是互补的，也就是说，如果其中一个假设为真，则另一个假设为假；如果我们推翻了其中一个假设，那就必须承认另一个假设。

第二，根据所收集到的数据，计算原假设成立下所得到的样本观察结果出现的概率。在统计学中，这个值被称为 p 值。我们姑且假定根据计算得到周三看电影的人数不比其他时间少在样本中出现的概率值 P=0.01。这样是不是就可以得出结论接受备择假设而否定原假设了呢？很显然，我们还缺少一个判断标准。

第三，确定判断标准，即选择恰当的小概率事件发生的概率值，我们用 a（Alpha）表示，统计学上称为显著性水平（Significance Level ）。通常用 a=0.05（5%）作为标准来进行决策：

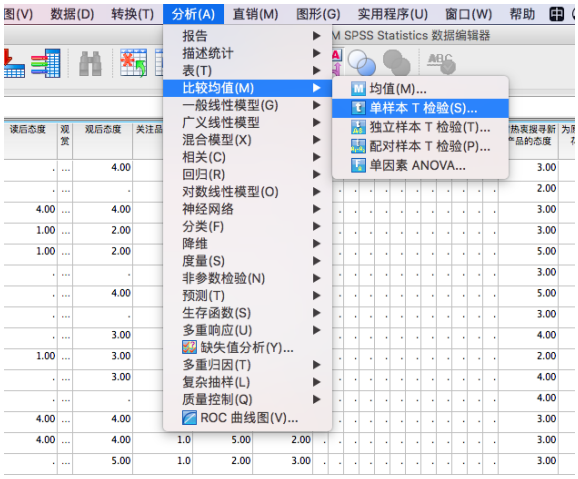
当 P≤a 时，拒绝原假设，接受备择假设

当 P＞a 时，接受原假设，拒绝备择假设

假设检验的过程同样是变量相关性分析的基本模式。

根据我们的计算，此时的 p=0.01＜a，所以根据 p 与 a 的关系可以得到周三看电影的人数比其他时间看电影的人数要少这一备择假设可被接受。

以上过程可以通过 SPSS 中单样本 T 检验计算得到：



在操作过程中，可以根据实际需要调整显著性水平 a（通常 a=0.1，0.05，0.01 等）。在输出结果中，显著性（双尾）即 Sig.（2-tailed）的值为我们所需要的 P 值。其中双尾代表原假设没有方向。如果原假设有方向，则因进行单尾检验，方法是将显著性（双尾）下的值除以 2 就可以了。（双尾检验和单尾检验可参见本章延伸与拓展部分）

小结：对单变量的描述可以为研究人员提供相关数据的详尽细节，通过可管理的方式呈现数据，保证了研究时的简单明了。如我们通过读均值、中位数和众数的观察，可以对变量的集中趋势有基本了解；通过对均值和标准差的观察，可以了解受访者的回答相较于均值的位置关系。

9.2.2 双变量统计分析

在实践中，研究人员不仅会对单一变量进行分析，也会对两个变量间的相互关系进行分析。以下我们将对常见的双变量统计分析方法进行介绍。

交叉表格

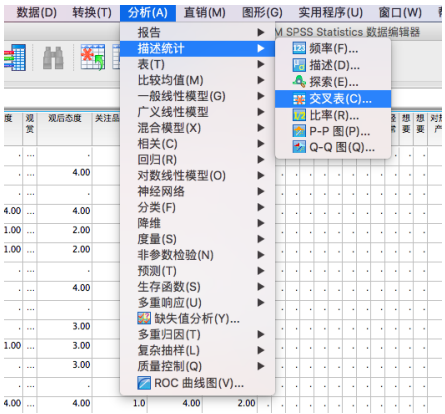
交叉表格（Cross Tabulation）是一种常用的分类汇总表，当需要分析的变量有一个是定类或两个都是定类变量时，可以使用交叉表格来清晰展示变量间的关系。图 9-5 是一个关于性别和出生年份的交叉表格。

1男2女 * 出生年份 交叉表

个数

		出生年份						
		1987.00	1988.00	1989.00	1990.00	1991.00	1992.00	總和
1男2女	1.00	1	1	1	5	2	1	11
	2.00	1	3	3	11	5	1	24
總和		2	4	4	16	7	2	35

同样，我们可以通过相关统计软件实现交叉表格。在 SPSS 中通过“分析”→“描述统计”→“交叉表”命令实现：



在“交叉表”弹出窗口，根据需要选择应进行交叉比较的两个变量分别至“行”、“列”后点击确定，就可以生成交叉表格。注意在“交叉表”弹窗中有“统计量”选项。此处点开后勾选相应统计方法（如卡方等），相关统计方法的意义我们会在后文进行介绍。这里只介绍通过 SPSS 实现交叉表格的方法。

卡方检验

研究人员经常需要对样本数据中两个变量之间的关系进行判断，当我们想知道样本在两个变量上的观测值是否有关联时，可以使用卡方分析、t 检验等。**卡方检验（Chi-square analysis，X^2）**适用于定类变量和定序变量。下面是一些可以通过卡方检验验证的例子：

看新闻的程度（低、中、高）与性别的关系

大学生和社会新鲜人（刚毕业参加工作 1-2 年）对于电影类型选择是否具有不同偏好

卡方检验逻辑与假设检验一致：

第一步 提出原假设（H_0）与备择假设（H_1）。通常假设两个变量没有关联。

第二步 计算相应统计值（此处为卡方值 X^2）

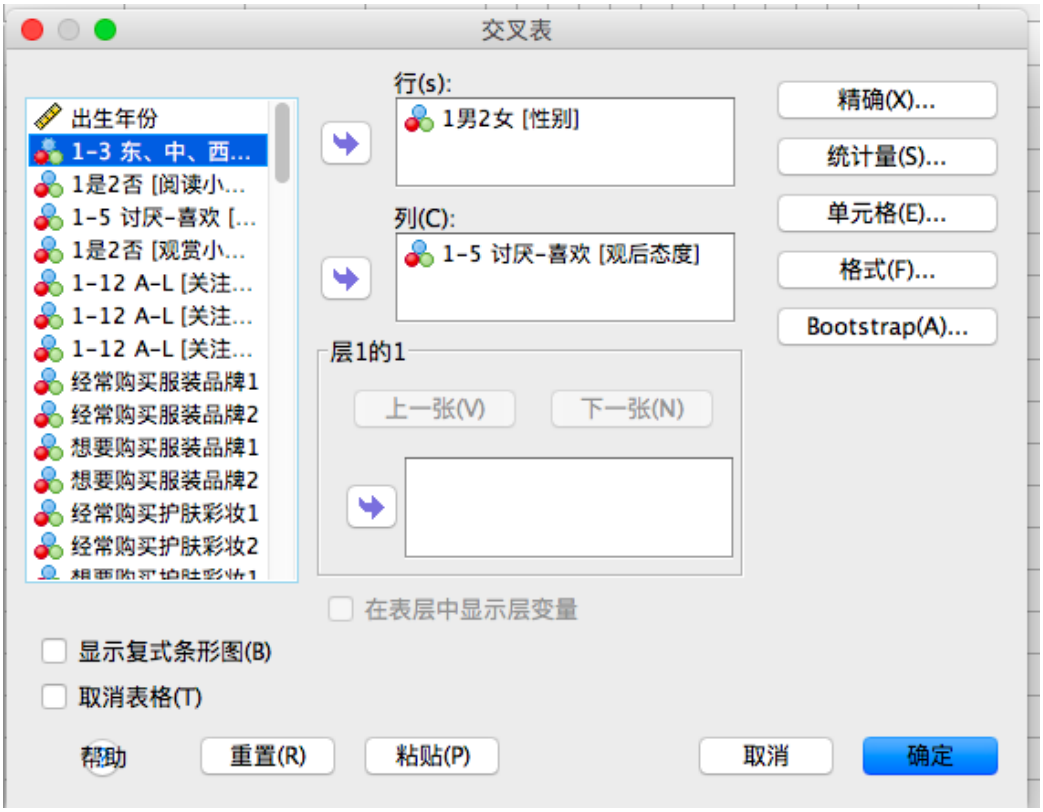
第三步 根据计算出的 X^2，求出原假设成立的概率值 P（可通过查询 X^2 分布表），并根据 P 与 a 的关系选择接受原假设或备择假设。

卡方值的计算公式为：

$$X^2 = \sum_{i=1}^{n} (\text{观测值}_i - \text{预期值}_i)^2 / \text{预期值}_i$$

在 SPSS 中可以通过交叉表格下“统计”选项实现卡方值的计算。具体操作过程如下：

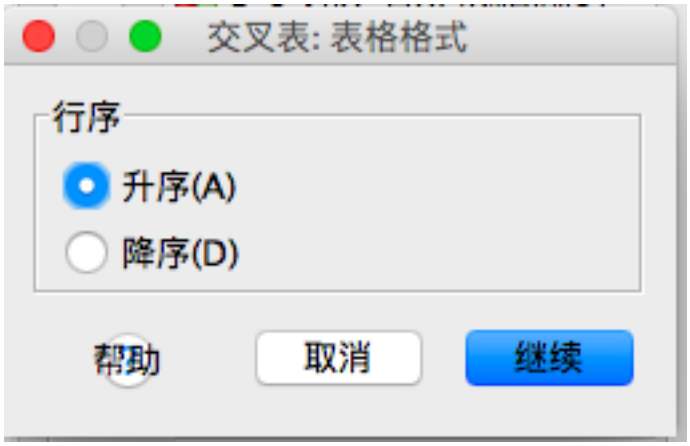
进入交叉表格，选择相应需要分析的变量进入行框和列框（可以根据实际需要选择一个额外的变量进入“层”框，这个变量将决定频数分布的层次；选择勾选“显示复式条形图”，则输出每个变量的分类条形图；勾选“取消表格”，则只输出统计量，而不输出交互分析表）



点击“统计量”选项，进入“交叉表：统计量”对话框，勾选卡方（卡方值只能证明两个变量的相关性，而无法证明相关性的强弱，此对话框中的其他选项（相关系数）可以对两个变量的相关性做出进一步判断）并点击继续。

点击“单元格”选项，对交叉表显示内容进行选择，并点击继续。

点击“格式”选项，选择对样本观察值进行升序或降序排列并点击继续。



最后点击确定，输出结果，下表中第一行第一列和第三列分别是 X^2 值和 P 值

卡方檢定

	數值	自由度	漸近顯著性(雙尾)
Pearson卡方			
概似比			
線性對線性的關連			
有效觀察值的個數			

在使用卡方检验时，需要注意两个问题。第一，卡方检验受样本量影响大，同样两个变量，不同的样本量，可能导出不同的结论。第二，对变量取值的不同分类会引起卡方值的改变，有可能得到不同的结果。

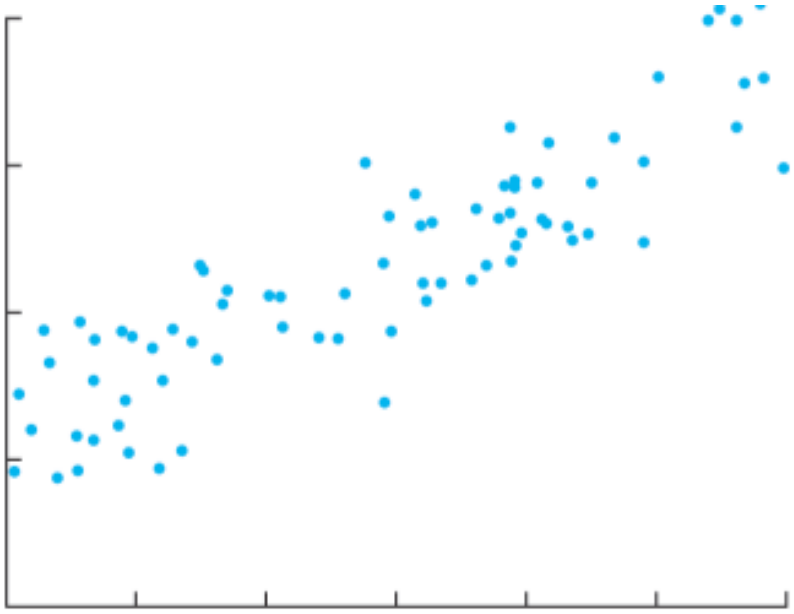
回归分析

回归分析（Regression analysis）目的在于了解在定距和定比变量中，一个变量的变化在多大程度上可以带来另一个变量的变化。更具体的来说，回归分析通过建立回归方程帮助人们了解在只有一个自变量变化时因变量的变化量。需要注意的是只有存在相关关系的变量才能进行回归分析。

（YouTube 扫码收看哈佛商业评论提供的对回归分析的解释 https://hbr.org/video/5299994733001/the-refresher-regression-analysis）

回归分析的理解相对复杂，以下通过一个假想的案例来帮助大家理解。

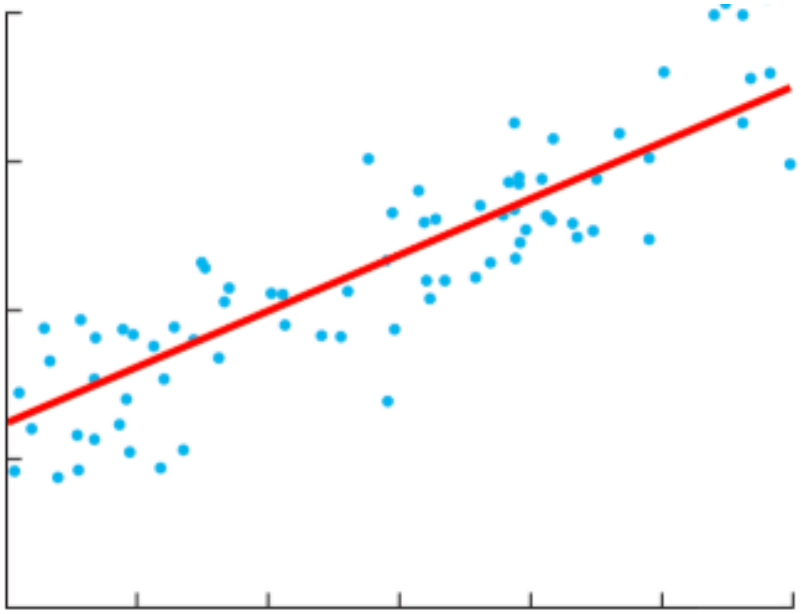
首先，假设我们已经有了关于一家小型电影院 3 年营业额的数据。我们感兴趣的是降雨量对营业额造成的影响有多大？在回归分析中，这些因素被称为变量。首先你需要有自己的因变量——你试图理解或预测的主要因素。在上面例子中，因变量是月营业额。然后是自变量，即你怀疑会对因变量有影响的变量，在这里是每月降雨量。如果我们把这些数据落到一张图上，它大体上应该是这样的：



其中 y 轴是营业额(因变量，请习惯将你感兴趣的东西放在 y 轴上)，x 轴是降雨量。每个点代表一个月的数据——包含当月的降雨量和销售额。

很显然，根据我们虚构的数据，雨天的时候该影院的营业额也相应更高。知道这个答案能够帮助我们确定这两个变量之间的确有关系，但这不是我们已经知道了的吗？（注意，我们已经提到了对两个变量进行回归分析的前提之一是这两个变量是相关的)所以我们想知道的是它们到底有多相关，即如果降雨量是 8cm（它可能不存在于所获得的现有数据中），是否能估算出对应月份影院的营业额是多少?如果降雨量是 10cm 呢?

我们可以想象在上面的图表中存在一条线（如下图所示），它大致贯穿所有数据点的中间。这条线将会在一定程度上帮助我们确定一般情况下雨天时影院的营业额。或许我们可以说这条线是对自变量和因变量之间关系的最佳解释。



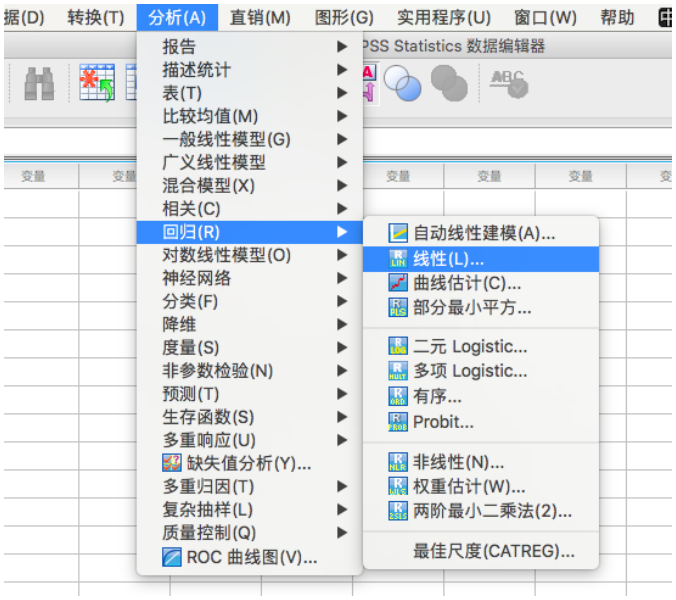
到这里，问题转变到了如何把这条想象出的线表达出来。我们需要建立一个关于自变量和因变量的方程式来表述这条最能代表自变量和因变量变化的线。根据示意图可以看出，我们的例子其实是最简单一元线性回归。一元线性回归的方程可以写成：

$$Y=a+bX$$

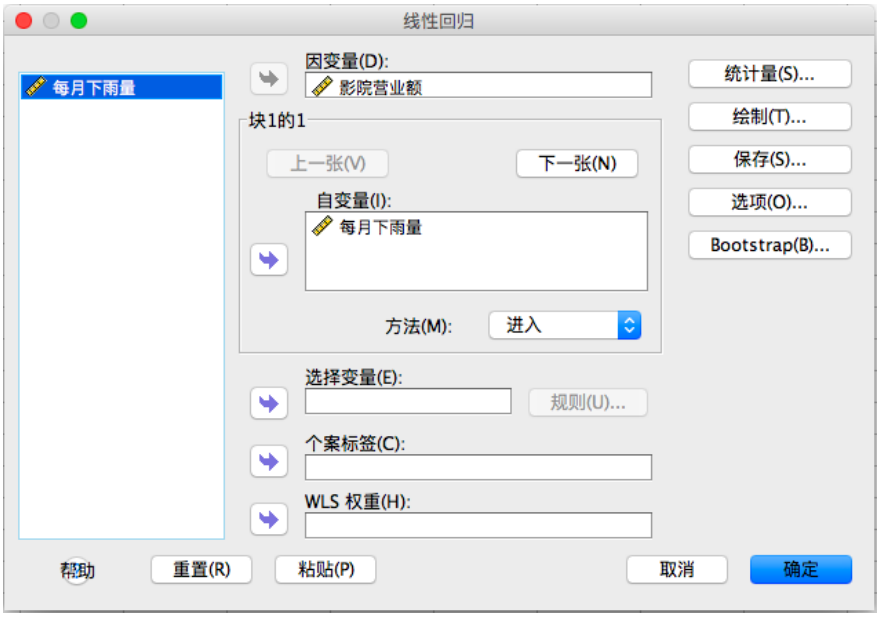
其中 X 代表自变量，Y 代表因变量。根据在中学所学的知识，a 代表直线（回归直线）的截距，b 代表回归直线的斜率，即回归系数。

通过 SPSS 可以很快得到 a 和 b 的值，进而有助于对 X 究竟怎么影响到 Y 的了解，其步骤如下：

“分析”→“回归”→“线性”



在弹出的“线性回归”窗口中，根据提示填入自变量 X 和因变量 Y，在本例中，我们真正感兴趣的是营业额，所以将其填入因变量。对营业额造成影响的降雨量填入自变量，并点击确定。



SPSS 将会自动输出结果，其中下面这张表上的数值就是我们想要的。

係數[a]

模式		未標準化係數 B 之估計值	未標準化係數 標準誤差	標準化係數 Beta 分配	t	顯著性
1	(常數)	479.075	35.525		13.486	.000
	每月下雨量	-5.502	1.712	-.713	-3.214	.009

a. 依變數: 影院营业额

根据结果，关于降雨量和影院营业额的一元线性回归方程可以写为：

Y=479.075+（-5.502）× X　（非标准化系数）

或

Y= -0.713 × X （标准化系数）

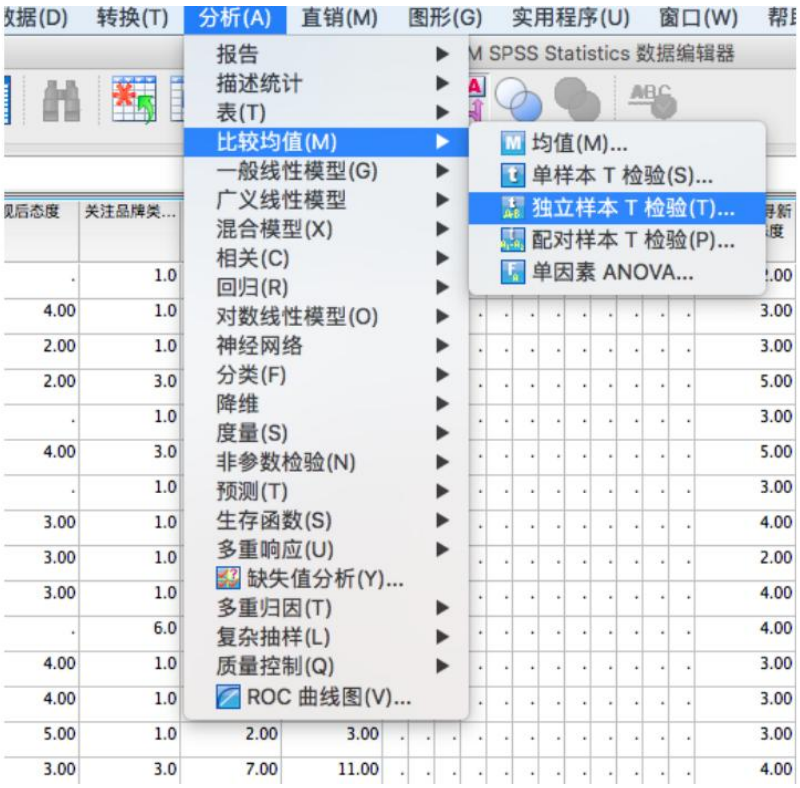
小结：线性回归是从一组样本数据出发，确定变量之间的数学关系式对这些关系式的可信程度进行各种统计检验，并从影响某一特定变量的诸多变量中找出哪些变量的影响显著，哪些不显著。利用所求的关系式，根据一个或几个变量的取值来预测或控制另一个特定变量的取值，并给出这种预测或控制的精确程度。在于通过后者的已知或设定值，去估计和（或）预测前者的（总体）均值。

变量间差异的比较：t 检验和方差检验

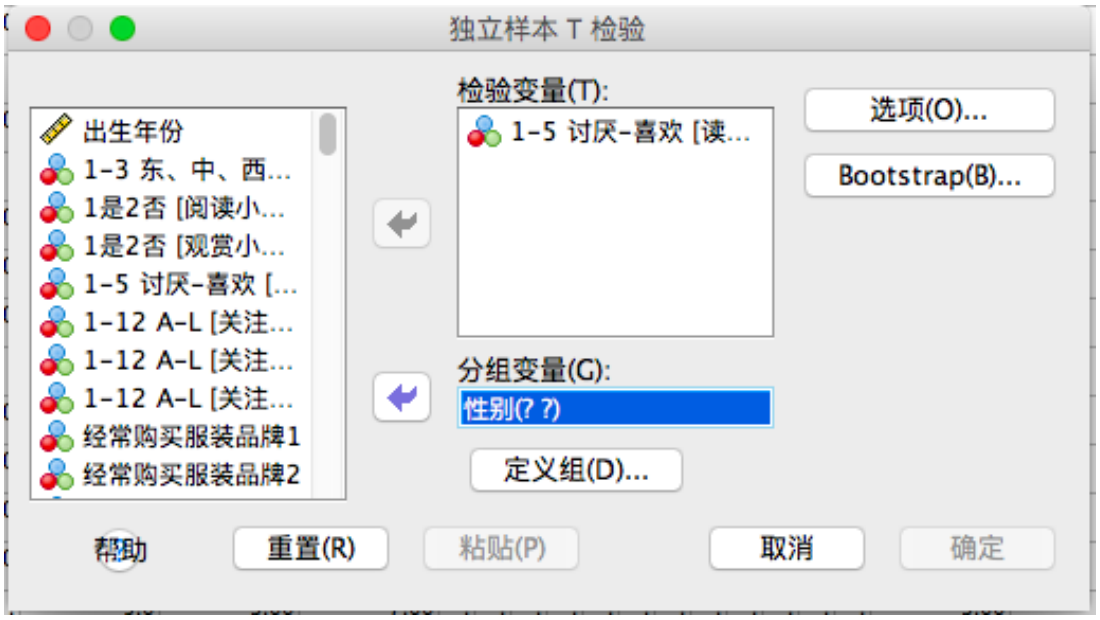
当自变量不仅是定类或定序变量，且该变量的属性类别只有两个，而因变量是定距或定比变量时，我们可以通过**独立样本 t 检验（Independent sample test）**来检验不同类别的两组自变量在因变量上的差异。比如男性和女性在观影时长上有没有差异（其对应的问题可以是（1）你的性别是 ___ （2）在观看电影时，你能接受的影片长度最长是 ___ 分钟）。进行 t 检验的前提假设是所检验的样本信息来源于一个服从正态分布的总体，且总体方差相等。

在 SPSS 中，依次点击“分析”→“比较均值”→“独立样本 T 检验”完成。具体步骤如下：

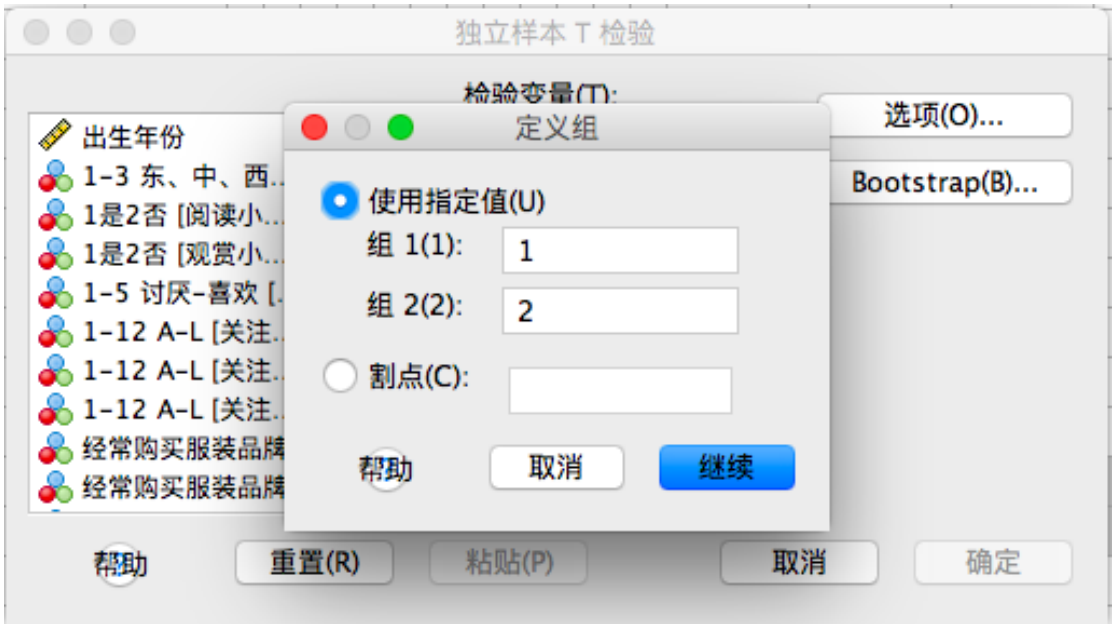
第一步，选择“独立样本 T 检验”



第二步，在“独立样本 T 检验”对话框中，将自变量选入“分组变量”，因变量选入“检验变量”



第三步，点击定义组，根据自变量编码确定“定义组”类别（示意图中，男性在录入时记为“1”，女性为“2”）



第四步，点击继续返回“独立样本 T 检验”，并点击确定。

以下是 SPSS 输出结果：

獨立樣本檢定

		變異數相等的 Levene 檢定		平均數相等的 t 檢定						
									差異的 95% 信賴區間	
		F 檢定	顯著性	t	自由度	顯著性 (雙尾)	平均差異	標準誤差異	下界	上界
1-5 讨厌-喜欢	假設變異數相等	.514	.488	-.298	11	.771	-.23333	.78309	-1.95691	1.49024
	不假設變異數相等			-.246	2.657	.823	-.23333	.94810	-3.48336	3.01669

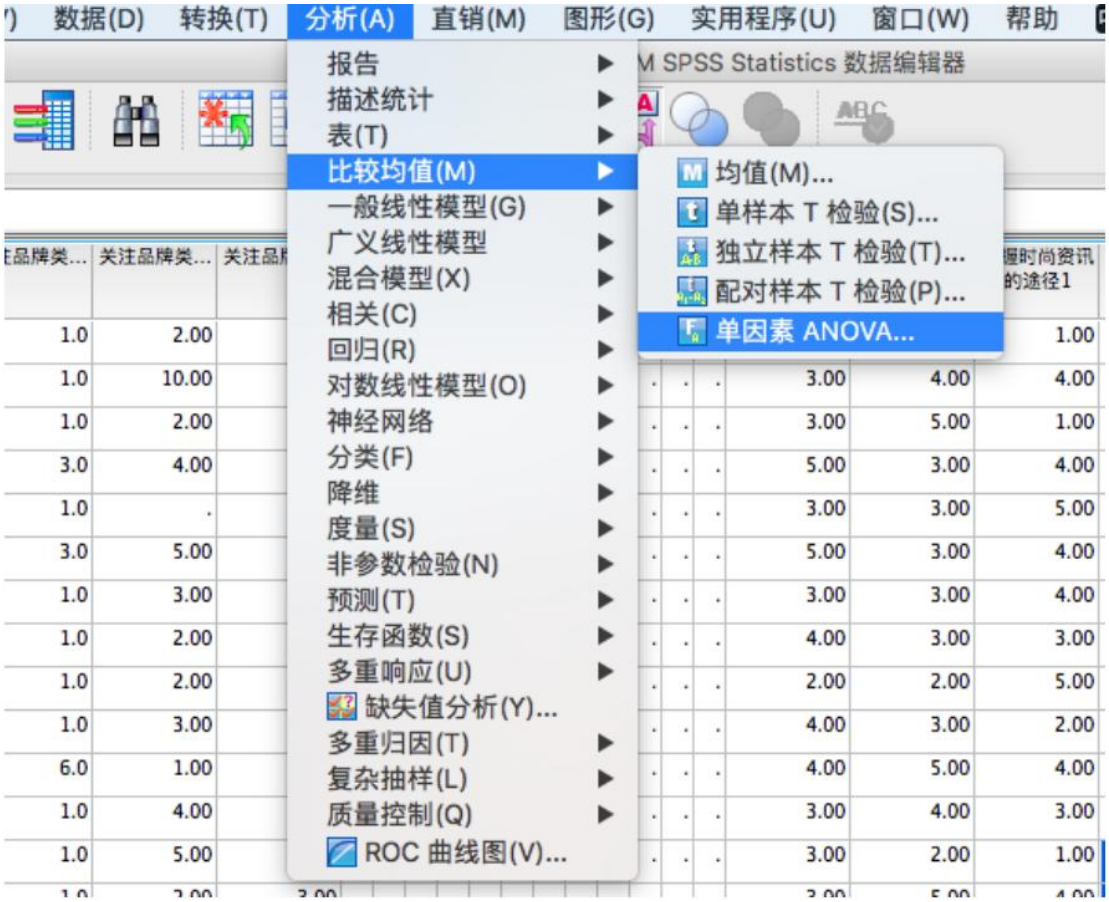
在上表中 Levene 检验是对方差齐性的检验。通过对 F 值所对应的 Sig（显著性）可以判断方差是否齐，如果 sig>0.05，说明满足方差齐性的条件，反之不满足。由于此时 Sig=0.488，因此方差是齐的。在“讨厌-喜欢”后给出了方差齐和不齐时的 t 检验结果。由于此时方差是齐的，所以我们只需要观察方差齐时所对应的数值即可。此时 t 检验的显著性（Sig（2-tailed））的值为 0.771，没有显著性，根据前文提到的一般步骤，此时因接受原假设，两个样本类别间没有差异。

方差检验

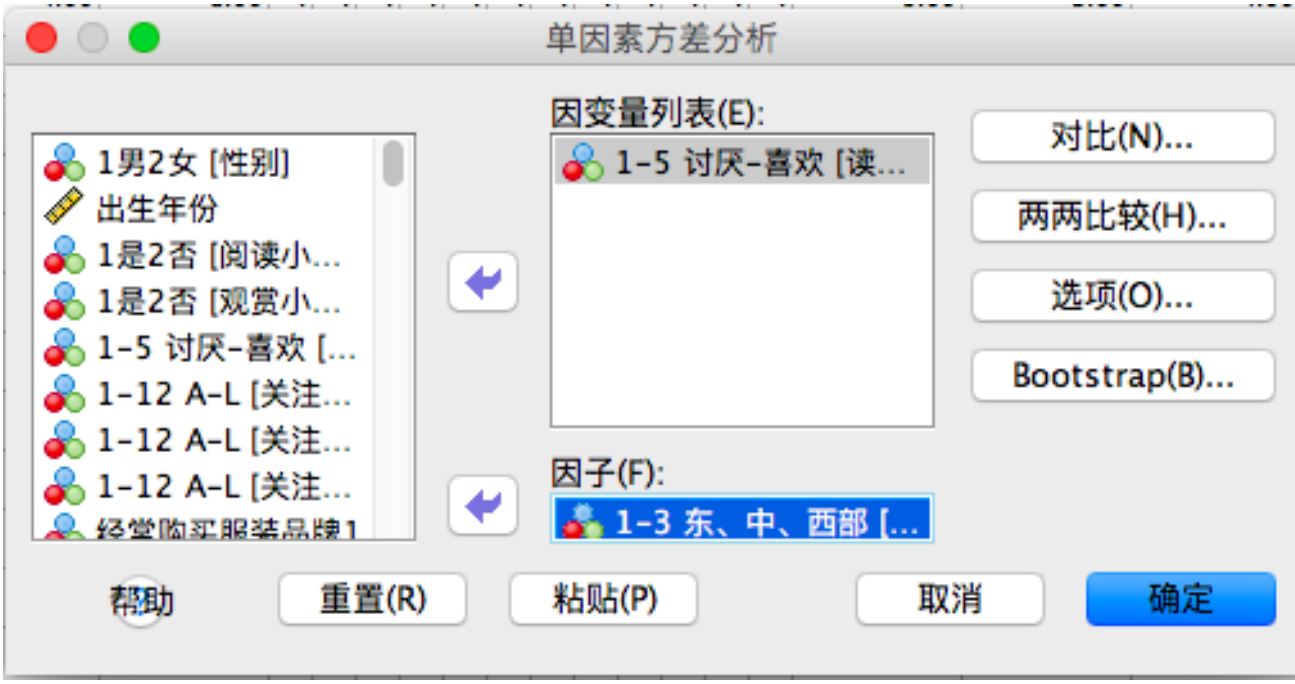
当自变量的类别不止两个，而是三个或以上（定类或定序）时，则可以通过**方差检验（Analysis of variance，ANOVA）**。比如我们想知道电影的不同类型和电影的豆瓣得分之间是否具有统计显著性，即类别上的差异在多大程度上影响了电影的得分。

通过 SPSS，我们可以轻松得到相关统计值，其步骤为“分析”→“比较均值”→“单因素 ANOVA”。

第一步，进入单因素 ANOVA



第二步，在弹出的“单因素方差分析”窗口中，将因变量填入“因变量列表”，将自变量填入“因子”，并点击确定。



其输出结果如下：

單因子變異數分析

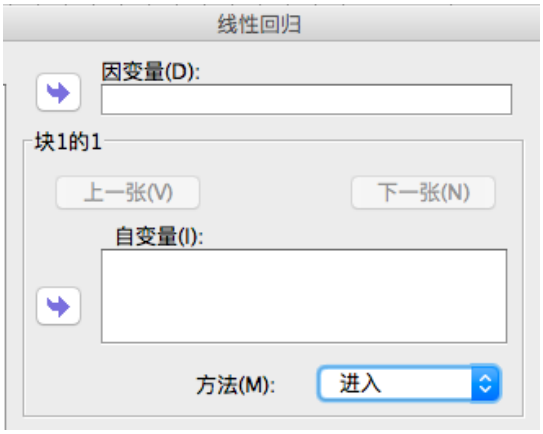
1-5 讨厌-喜欢

	平方和	自由度	平均平方和	F	顯著性
組間	2.942	2	1.471	1.154	.354
組內	12.750	10	1.275		
總和	15.692	12			

从结果可以看出，此时 F 值所对应的概值（显著性，sig）为 0.354，大于 0.05，接受原假设即不同类别在得分上没有差异。

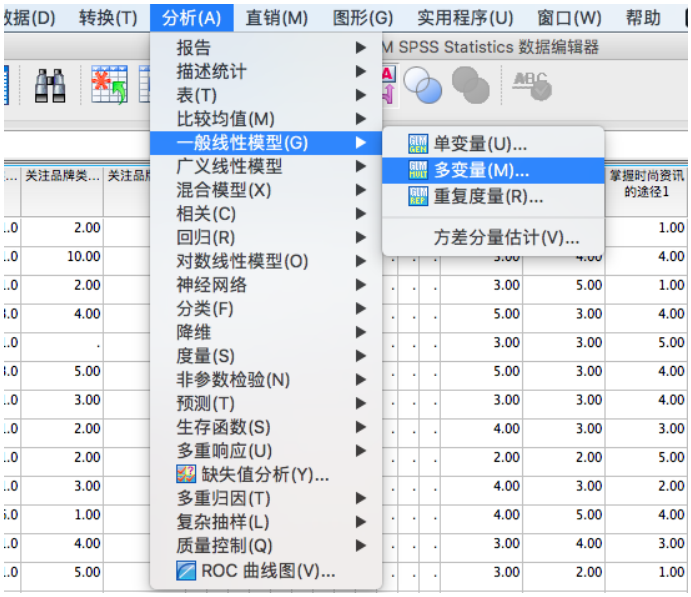
9.2.3 多变量统计分析

当我们想通过对多个自变量的共同变化的观察来预测一个因变量的变化时，一元线性回归就无法满足我们的需求了。此时，我们需要运用到多元线性回归。在 SPSS 中的操作方法与一元线性回归相同，即通过“分析”→“回归”→“线性”打开对话框进行操作。通过对结果显著性的判断回答研究问题。和一元线性回归操作的不同在于，多元线性回归中，自变量的选择不再是一个，而是多个。需要注意的是多元线性回归要求样本量和自变量个数的比值大于等于 5，且如果在方法上选择“逐步”，则这个比值因在 50 以上。



在方差分析中，如果因变量不是一个，而是两个（定距或定比）以上时，在 SPSS 中则不能使用单因素方差分析，而是应该使用多元方差分析（MANOVA），其操作步骤为：“分析”→“一般线性模型”→“多变量”。

首先，进入“多变量”窗口。



然后，将所需分析的变量（因变量）选入“因变量”窗口，将自变量（分组变量）选入“固定因子”窗口，单击确定即可。

根据 SPSS 的输出结果，通过对显著性的检测给出结论。其逻辑与单因素方差分析一样，不再赘述。

此外在多变量分析中，还有一种简化量表结构的检验方式——因子检验。通过对一份量表（自变量）中不同维度的描述进行检验，可以判断这些描述（因子）是否能够有效的测量出因变量。在大规模调研中，研究人员往往通过因子分析来简化数据。

在 SPSS 中，因子分析可以通过“分析”→“降维”→“因子分析”运算实现。根据 KMO 与 Bartlett’s 检测结果表中 MSA（取样适切性量数）来判断因子分析条件是否满足，一般 MSA 值在 0.5-1 间时，我们认为可以接受因子分析结果，小于 0.5 则不接受因子分析结果，

停止因子分析。当 MSA=1 时，意味着一个变量可以通过其他变量来解释。只有通过 KMO 与 Bartlett’s 检测，我们才能进行后续的分析。

需要注意的是，严格意义上只有定距和定比测量才适用于因子检验。但在实际运用中，很多量表都是定序量表，所以部分研究人员为了便于研究，也将其视为定距量表来对待。

由此我们可以扩展总结出针对多变量的分析方法，它们包括上述已经提到的多元方差分析、因子分析，以及主成分分析、典型相关、聚类分析、判别分析等，在此就不再进一步展开说明了。