

第 11 章 描述性统计分析

——Descriptive Statistics 菜单详解


知己知彼，百战不殆

——孙子

统计分析的目的是研究总体特征。但是，由于各种各样的原因，我们能够得到的往往只能是从总体中随机抽取的一部分观察对象，它们构成了样本，只有通过对样本的研究，我们才能对总体的实际情况做出可能的推断。因此，描述性统计分析是统计分析的第一步，做好这第一步是下面进行正确统计推断的先决条件。

依照 SPSS 的分类方式，我们平时所用到的描述统计量主要有以下几类：

- ✧ 集中趋势指标(Central Tendency)：均数、众数、中位数、几何均数、调和均数、总和。其中均数适用于正态分布和对称分布资料，中位数则适用于所有分布类型的资料。

 众数(Mode)指所有数值中出现频率最高的一个值，在国内用的非常少。几何均数和调和均数则不能用本章讲解的模块直接求出，但可以用 OLAP 和 Means 过程求得。

- ✧ 离散趋势指标(Dispersion)：标准差、方差、全距、最小值、最大值、标准误。其中标准差、方差只适用于正态分布资料，标准误则实际上反映了样本均数的波动程度。
- ✧ 百分位数指标(Percentile)：包括四分位数、各个百分位数等，适用于任何分布类型资料。
- ✧ 分布指标(Distribution)：偏度系数、峰度系数，它们反映了数据偏离正态分布的程度。
- ✧ 其他：M 统计量(M-estimators)、极端值(Outlier)等，它们主要用于对存在异常值的数据进行描述。

所用到的统计图则有：条图、饼图、直方图、箱式图、QQ 图（用于判断正态性）等。

SPSS 的许多模块均可完成描述性分析，但专门为该目的而设计的几个模块则集中在 Descriptive Statistics 菜单（见图 11.1）中，它们就是计算上面提到的各种统计量或绘出统计图来实现描述功能。这些模块是：

- ✧ Frequencies 过程：其特色是产生频数表，对分类资料和定量资料都适用。
- ✧ Descriptives 过程：进行一般性的统计描述，适用于服从正态分布的定量资料。该过程的功能比较简单，但使用频率却是最高的。
- ✧ Explore 过程：顾名思义，该过程用于对数据分布状况不清时的探索性分析，它会杂七杂八地给出一大堆可能用到的统计指标和统计图，好让研究者在大海里捞针。

- ◇ Crosstabs 过程：完成分类资料/等级资料的统计描述和各种各样“常规”的统计检验，常用的 χ^2 检验也在其中完成。
- ◇ Ratio 过程：是 SPSS 11.0 版新增的方法，用于对两个连续性变量计算相对比指标，它可以计算出一系列非常专业的相对比描述指标，其中的大多数还为我们所不太熟悉。




图 11.1 Descriptive 菜单

由于 Crosstabs 过程不仅仅限于对资料的描述，而涉及到了比较详尽的分类资料统计分析方法，我们将对其另辟专章讲解，本章只学习另四个过程。

 这些过程在 9.0 及以前版本中被放置在 Summarize 菜单中。

11.1 Frequencies 过程

频数分布表是描述性统计中最常用的方法之一，Frequencies 过程就是专门为产生频数表而设计的。它不仅产生详细的频数表，还可以按要求给出某百分位点的数值，以及常用的条图、圆图等统计图。相比之下，它更适合于对分类变量以及不服从正态分布的连续性变量进行描述。

 和国内常用的频数表不同，几乎所有统计软件给出的均是详细频数表，即并不按某种要求确定组段数和组距，而是按照数值精确列表。如果想用 Frequencies 过程得到我们所熟悉的频数表，请先用第 3 章学过的 Recode 过程产生一个新变量来代表所需的各组段。

11.1.1 引例

例 11.1 某地 101 例健康男子血清总胆固醇值测定结果如下，请绘制频数表、直方图，计算均数、标准差、变异系数 CV、中位数 M、p2.5 和 p97.5（杨树勤，《卫生统计学》第三版 233 页）。

4.77	3.37	6.14	3.95	3.56	4.23	4.31	4.71	5.69	4.12	4.56	4.37	5.39	6.30	5.21
7.22	5.54	3.93	5.21	4.12	5.18	5.77	4.79	5.12	5.20	5.10	4.70	4.74	3.50	4.69
4.38	4.89	6.25	5.32	4.50	4.63	3.61	4.44	4.43	4.25	4.03	5.85	4.09	3.35	4.08
4.79	5.30	4.97	3.18	3.97	5.16	5.10	5.86	4.79	5.34	4.24	4.32	4.77	6.36	6.38
4.88	5.55	3.04	4.55	3.35	4.87	4.17	5.85	5.16	5.09	4.52	4.38	4.31	4.58	5.72
6.55	4.76	4.61	4.17	4.03	4.47	3.40	3.91	2.70	4.60	4.09	5.96	5.48	4.40	4.55
5.38	3.89	4.60	4.47	3.64	4.34	5.18	6.14	3.24	4.90	3.05				

解：数据已录入为文件 dguchun.sav，变量名为 X。显然，血清总胆固醇值是一个定量指标，对定量指标的统计描述就应当用相应的三个过程来完成，但此处要求绘制频数表，显然是 Frequencies 过程的特长，同时 P2.5 和 P97.5 这两个特殊的百分位数也

只有它能够求出。题中要求的变异系数是无法直接得到的，可以用均数和标准差手工计算。操作如下：

Analyze→Descriptive Statistics→Frequencies

Variables 框: X

选入要分析的变量

Statistics:

☒ Mean, ☒ Median, ☒ Std. deviation

要求计算均数、标准差和中位数

☒ Percentiles: 键入 2.5:

要求计算 P2.5 和 P97.5 百分位数

☒ Percentile: 键入 97.5:

Charts:

☒ Bar charts

做出频数分布的直方图(条图)

得出结果后手工计算出 CV。

 上面做出的直方图分组太多，需要进一步编辑，详情请参见绘图章节。

11.1.2 界面说明

在引例中我们用到的 Frequencies 对话框界面如图 11.2 所示。

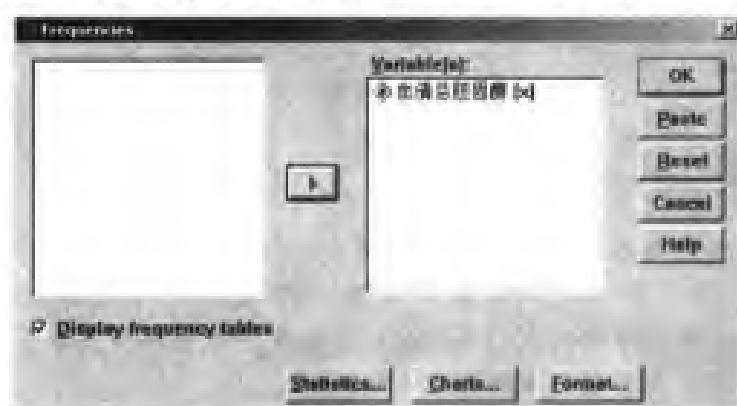


图 11.2 主对话框

该界面在 SPSS 中实在太普通了，无须多言，重点介绍一下各部分的功能：

【主对话框】

1. Variable(s)框：用于选入需要进行描述的变量，如果选入多个，系统会对其依次进行分析。

2. ☒ Display frequency tables：确定是否在结果中输出频数表。

【Statistics子对话框】（见图 11.3）

该对话框的功能为定义需要计算的其他描述统计量。

1. Percentile Values 复选框组：定义需要输出的百分位数，可计算四分位数(Quartiles)、每隔指定百分位输出当前百分位数(Cut points for equal groups)或直接指定某个百分位数(Percentiles)，本例就在其中直接指定输出 P2.5 和 P97.5。

2. Central tendency 复选框组：用于定义描述集中趋势的一组指标：均数(Mean)、

中位数(Median)、众数(Mode)、总和(Sum)。

3. Dispersion 复选框组：用于定义描述离散趋势的一组指标：标准差(Std. Deviation)、方差(Variance)、全距(Range)、最小值(Minimum)、最大值(Maximum)、标准误(S.E. mean)。

4. Distribution 复选框组：用于定义描述分布特征的两个指标：偏度系数(Skewness)和峰度系数(Kurtosis)。

5. ☐ Values are group midpoints：当输出的数据是分组频数数据，并且具体数值是组中值时，选中该复选框以通知 SPSS，这样它在计算各种百分位数的时候会将数据按频数表对待，而不会认为同一组内的数据取值都是组中值的大小。当然，如果你不计算百分位数，选不选它无所谓。

【Charts】子对话框（见图 11.4）

该对话框用于设定所做的统计图。

1. Chart type 单按钮组：定义统计图类型，有四种选择：无、条图(Bar charts)、圆图(Pie charts)、直方图(Histograms)。其中直方图还可以选择是否加上正态曲线(With normal curve)。这里我们需要选择绘制直方图。

2. Chart Values 单按钮组：当选择绘制条图和圆图时定义是按照频数还是按百分比做图（即影响纵坐标刻度）。

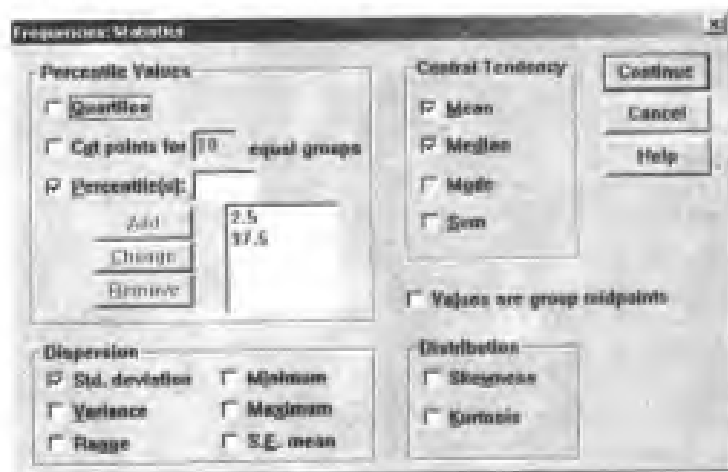


图 11.3 Statistics 子对话框

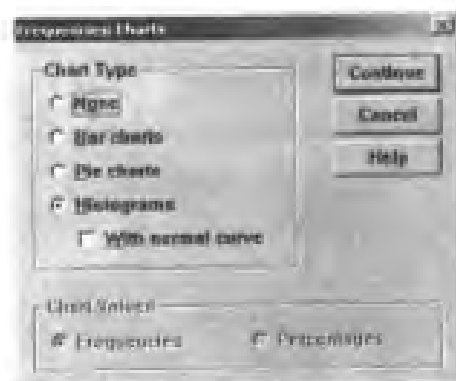


图 11.4 Charts 子对话框

【Format】子对话框（见图 11.5）

用于定义输出频数表的格式，不过一般不用更改，使用默认设置即可。

1. Order by 单按钮组：定义频数表的排列次序，有四个选项：

- ✧ Ascending values：根据数值按升序从小到大作频数分布。
- ✧ Descending values：根据数值按降序从大到小作频数分布。
- ✧ Ascending counts：根据频数按升序从少到多作频数分布。
- ✧ Descending counts：根据频数按降序从多到少作频数分布。

2. Multiple Variables 单按钮组：如果选择了两个以上变量做频数表，则 Compare variables 可以将它们的结果在同一个频数表过程输出结果中显示，便于互相比较：

Organize output by variables 则将结果在不同的频数表过程输出结果中显示。

3. ☐ Suppress tables more than...: 当频数表的分组数大于下面设定数值时禁止它在结果中输出, 这样可以避免产生巨型表格。



图 11.5 **Format** 子对话框

11.1.3 结果解释

引例的输出结果如下:

Frequencies

Statistics		
血清总胆固醇		
N	Valid	101
	Missing	0
Mean		4.6995
Median		4.6100
Std. Deviation		.8616
Percentiles	2.5	3.0455
	97.5	6.4565

最上方为表格名称, 左上方为分析变量名, 可见样本量 N 为 101 例, 缺失值 0 例, 均数 $\text{Mean}=4.69$, 中位数 $\text{Median}=4.61$, 标准差 $\text{STD}=0.8616$, $P_{2.5}=3.04$, $P_{97.5}=6.45$ 。我们需要的统计指标基本上都在这里了。

根据上述结果, 我们就可以计算出变异系数 $\text{CV}=0.8616/4.69=0.1837$ 。

血清总胆固醇

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2.70	1	1.0	1.0	1.0
	3.04	1	1.0	1.0	2.0
	3.05	1	1.0	1.0	3.0
	3.18	1	1.0	1.0	4.0
	3.24	1	1.0	1.0	5.0
	3.35	2	2.0	2.0	6.9

系统对变量 x 作频数分布表 (此处只列出了开头部分), Valid 右侧为原始值, Frequency 为频数, Percent 为各组频数占总例数的百分比 (包括缺失记录在内), Valid percent 为各组频数占总例数的有效百分比, Cum Percent 为各组频数占总例数的累积百分比。

在图 11.6 中，左图即为绘制出的直方图，可见数据基本上呈正态分布。右侧的图例中会给出均数和标准差。如果希望和理论上的正态分布曲线作比较，则可以进入编辑状态，在 Options 中选择显示正态曲线，结果如右图所示。可见分布是和正态曲线比较吻合的。

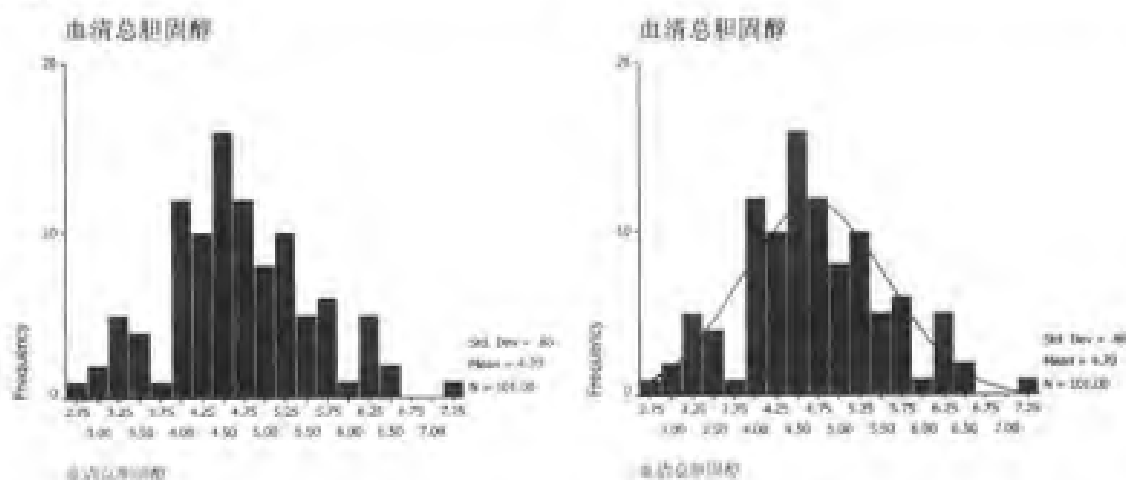


图 11.6 原始直方图和添加正态曲线后的直方图

11.2 Descriptives 过程

Descriptives 过程是连续资料统计描述应用最多的一个过程，它可对变量进行描述性统计分析，计算并列出一系列相应的统计指标，这和其他过程相比并无不同。但该过程还有个特殊功能，那就是可将原始数据转换成标准正态评分值，并以变量的形式存入数据库供以后分析。

11.2.1 界面说明

【主对话框】（见图 11.7）

1. Variable(s)框：用于选入需要进行描述的变量，如果选入多个，系统会对其依次进行描述，但输出在同一张表格内。

2. ☐ Save standardized values as variables: 确定是否将原始数据的标准正态变换结果存为新变量，选中它会在数据集中生成一个新的变量，该变量自动命名为“Z+原变量名”，大小即为原变量的标准正态变换结果。

【Options】子对话框（见图 11.8）

大部分内容均在前面 Frequencies 过程的 Statistics 对话框中见过，只有最下方的 Display Order 单选按钮组是新的，可以选择为变量列表顺序、字母顺序、均数升序或均数降序。



图 11.7 主对话框

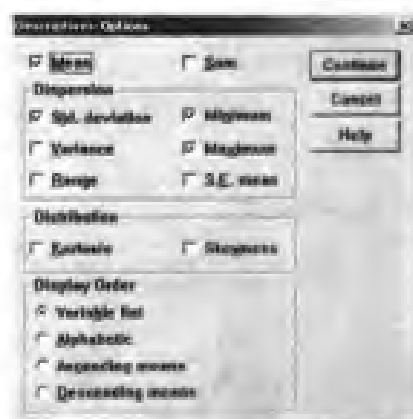


图 11.8 Options 子对话框

11.2.2 结果解释

下面是按照默认选项对例 11.1 使用 Descriptives 过程进行分析的结果输出：

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
血清总胆固醇	101	2.70	7.22	4.6995	.8816
Valid N (listwise)	101				

一望可知，这里的大部分内容都在上一节见过，因此就不再多解释了。

讲了两个过程，也许大家已经发现：多数结果中的统计专业单词在对话框中就已经出现，因此我们以后会详细解释对话框的内容，结果中相同的单词不再重复解释。

11.3 Explore 过程

Explore 过程是以上三个过程中功能最为强大的一个，它可对变量进行更为深入详尽的描述性统计分析，主要用于对资料的性质、分布特点等完全不清楚时，故又称之为探索性分析。在常用描述性统计指标的基础上，它又增加了有关数据详细分布特征的文字与图形描述，如茎叶图、箱式图等，显得更加详细、全面。还可以为以方差齐性为目的的变量变换提供线索，有助于用户制定继续分析的方案。

11.3.1 界面说明

【主对话框】（见图 11.9）

1. Dependent List 框：用于选入需要分析的变量。
2. Factor List 框：如果想让所分析的变量按某种因素取值分组分析，则在这里选入分组变量。

3. Label Cases by 框：选择一个变量，它的取值将作为每条记录的标签。最典型的情况是使用记录 ID 号的变量。

4. Display 单选按钮组：用于选择结果中是否包含统计描述、统计图或两者均包括。

【Statistics 子对话框】（见图 11.10）

用于选择所需要的描述统计量，有如下选项：

1. ☒ Descriptives：输出均数、中位数、众数、5%修正均数、标准误、方差、标准差、最小值、最大值、全距、四分位全距、峰度系数、峰度系数的标准误、偏度系数、偏度系数的标准误及指定的均数可信区间。

2. ☐ M-estimators：作集中趋势的最大稳健估计，该统计量是利用迭代方法计算出来，一般来说受异常值的影响要小的多。如果该估计量离均数和中位数较远，则说明数据中可能存在异常值，此时宜用该估计值替代均数以反映集中趋势。一共会输出 Huber、Andrew、Hampel 和 Tukey 四种 M 统计量，其中 Huber 法适用于数据接近正态分布的情况，另三种则适用于数据中有许多异常值时。

3. ☐ Outliers：输出五个最大值与五个最小值。

4. ☐ Percentiles：输出第 5%、10%、25%、50%、75%、90%、95% 分位数。

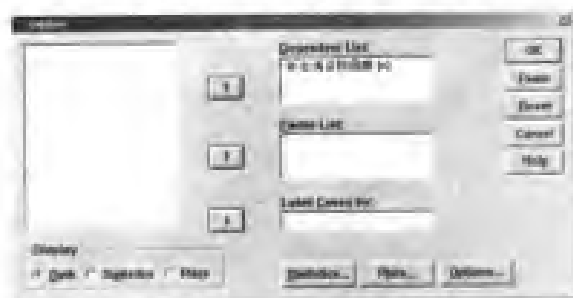


图 11.9 主对话框



图 11.10 Statistics 子对话框

【Plots 子对话框】（见图 11.11）

用于选择所需要的统计图，有如下选项：

1. Boxplots 单选按钮组：确定箱式图的绘制方式，可以是按组别分组绘制(Factor levels together)，也可以不分组一起绘制(Dependents together)，或者不绘制(None)。

2. Descriptive 复选按钮组：可以选择绘制茎叶图(Stem-and-leaf)和直方图(Histogram)。

3. ☐ Normality plots with test：绘制正态分布图，并进行变量是否符合正态分布的检验。

4. Spread vs. Level with Levene Test 单选按钮组：该部分属于高级分析功能，当选入分组变量时可用，其目的是判断各组间的离散程度是否相同，并为此寻求一个比较合适的变量变换方法。具体会输出分布——水平图，给出回归直线斜率，并进行稳健的 Levene 方差齐性检验。

◇ None：什么都不做，系统默认。

◇ Power estimation：用于帮助估计对原始数据应当进行指数为多少的幂函数

($y=x^p$) 变换才能使得各组间的方差最齐。所做出的散点图横轴为各组中位数的自然对数, 纵轴为各组四分位数间距的自然对数。图形下方会给出相应直线的斜率和最佳转换幂次的估计值。

- ✧ **Transformed:** 提供了几种常用的幂函数变换方法, 输出的散点图将按照变换后的数据来绘制, 横轴是变换后的中位数, 纵轴是变换后的四分位间距。具体的幂次在右侧 **Power** 下拉列表中选择, 可以使用的有自然对数、-1/2 次方 (1/Square root)、-1 次方 (Reciprocal)、1/2 次方 (Square root)、平方 (Square)、三次方 (Cube)。此处应当根据上面 **Power estimation** 的结果选择最接近的一个幂次。
- ✧ **Untransformed:** 不对数据进行转换, 直接使用原始数据绘图, 这相当于幂次为 1 的变换。

【Options 子对话框】(见图 11.12)

用于选择对缺失值的处理方式, 可以是不分析有任一缺失值的记录、不分析计算某统计量时有缺失值的记录, 或报告缺失值。

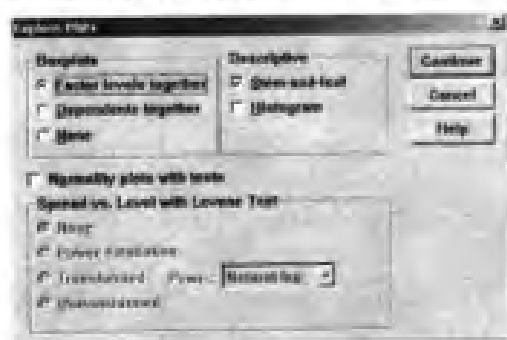


图 11.11 Plots 子对话框



图 11.12 Options 子对话框

11.3.2 结果解释

以例 11.1 的数据为例, 按默认方式下的选择, Explore 过程的输出如下:

Explore

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
血清总胆固醇	101	100.0%	0	0%	101	100.0%

首先是例行的处理记录缺失值情况报告, 可见 101 例均为有效值。

Descriptives				
			Statistic	Std. Error
血清总胆固醇	Mean		4.6995	8.573E-02
	95% Confidence Interval for Mean	Lower Bound	4.5294	
		Upper Bound	4.8696	
	5% Trimmed Mean		4.6881	
	Median		4.6100	
	Variance		.742	
	Std. Deviation		.8616	
	Minimum		2.70	
	Maximum		7.22	
	Range		4.52	
	Interquartile Range		1.0600	
	Skewness		.251	.240
	Kurtosis		.101	.476

上表详细列出了常用的描述统计量，如果有标准误也会列出（如偏度和峰度系数）。其输出内容中的统计量从上至下依次为：均数（及标准误）、均数 95%可信区间下限、上限、去除 5%极端值后的均数、中位数、方差、标准差、最小值、最大值、全距、四分位数间距、偏度系数（及标准误）、峰度系数（及标准误）。

图 11.13 分别是茎叶图和箱式图。茎叶图的排列方式和频数表有些类似，不过改成了整数位合在一起，称为茎，图的下方会标示出茎宽和实际值的倍数，如茎宽为 10，则图中的 2.3 代表实际取值 23。本图中为 1，则图中取值就是实际值；将小数位单独列出，称为叶，同样在图的下方会标示出每片叶子代表几个实际记录，本图中仍为 1。茎叶图可以非常直观的看出数据的分布范围及形态，在国外非常流行。

血清总胆固醇

血清总胆固醇 Stem-and-Leaf Plot

Frequency	Stem &	Leaf
1.00	2 .	7
8.00	3 .	00123334
9.00	3 .	556689999
24.00	4 .	00000111122233333334444
25.00	4 .	555555666667777777788899
17.00	5 .	01111111222333334
9.00	5 .	556778889
8.00	6 .	112333
1.00	6 .	5
1.00	Extremes	(>=7.2)
Stem width: 1.00		
Each leaf: 1 case(s)		

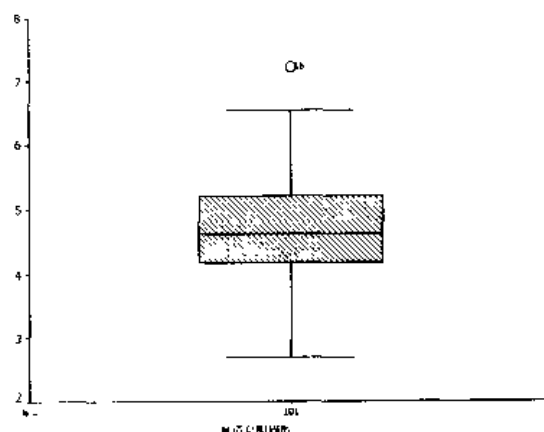


图 11.13 茎叶图和箱式图

在箱式图里，中间的黑粗线为均数，方框为四分位间距的范围，上下两个细线为除了离群值之外的最大、最小值，这两个细线之外的数据点均为可疑的离群值

(Outlier) 或极值 (Extreme Value), 其中超过四分位数间距 1.5 倍的为离群值, 以 “O” 表示; 超过 3 倍的则为极值。以 “*” 表示, 如本图中就有一个离群值, 旁边标注的是该数据点的记录号 15, 提示我们要对该记录加以关心。但总的来说, 数据的分布是相当对称的。

11.3.3 对引例的进一步分析

【M 统计量】

如果在 **Statistics** 子对话框中选中 M-estimators, 则结果会输出 M 统计量如下:

M-Estimators				
	Huber's M-Estimator ^a	Tukey's Biweight ^b	Hampel's M-Estimator ^c	Andrews' Wave ^d
血清总胆固醇	4.6640	4.6442	4.6711	4.6443

^a The weighting constant is 1.339.

^b The weighting constant is 4.685

^c The weighting constants are 1.700, 3.400, and 8.500

^d The weighting constant is 1.340*pt.

可见 Huber、Andrew、Hampel 和 Tukey 四种 M 统计量的数值非常接近, 且离计算出的均数值都不太远。这部分说明数据的分布不太偏, 均数是可以代表数据的集中趋势的。

【正态概率图】

利用 **Plot** 子对话框中的 Normality plots with test 复选框, 可做出 QQ 正态概率图和去势 QQ 正态概率图如图 11.4 所示。

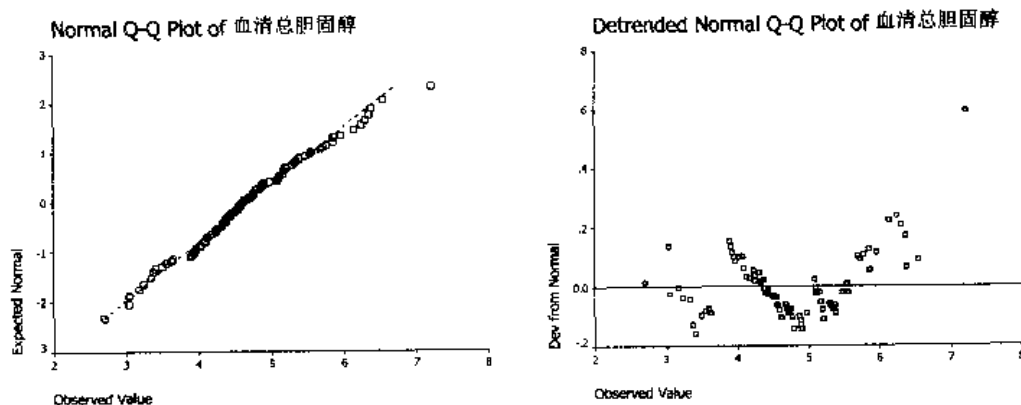


图 11.14 QQ 正态概率图和去势 QQ 正态概率图

左侧为正态 QQ 图, 如果数据呈正态分布, 则图中数据点应和理论直线基本重合。可见数据基本还是在直线上的, 只是最大的那个数据较为偏离。但从总体而言, 数据并未出现明显违反正态分布的情况。

为了更仔细的观察, 我们可以看右侧的趋势 QQ 图, 该图反映的是按正态分布计算的理论值和实际值之差的分布情况。如果数据服从正态分布, 则数据点应较均匀的分布在 $Y=0$ 这条直线上下。图中可见最大值那个点的确离理论分布线较远, 这和前面箱式图的结果一致。

11.4 Ratio 过程

Ratio 过程是 SPSS 11.0 版新增的方法，用于对两个连续性变量计算相对比指标，当研究者关心 A、B 两个指标比值的变动情况时，该过程非常有用。它可以计算出一系列非常专业的相对比描述指标，其中的大多数还为我们所不太熟悉。

11.4.1 引例与界面说明

例 11.2 对 SPSS 自带数据集 cars.sav 中的汽车功率和车重之比进行描述，并观察不同产地的汽车该指标是否有差异。

解：本题要进行描述的是汽车功率和车重之比，有两种方法可以实现，第一种是首先计算出一个新变量代表每辆车的比值，然后对该变量进行描述；第二种方法是利用 Ratio 过程加以分析，由于后者可以计算出一些专业的指标，显然要更合适一些，这里演示其操作如下：

Analyze→Descriptive Statistics→Ratio

Numerator 框：horse

选入要分析的变量

Denominator 框：weight

Sort by group variable 框：origin

Statistics:

☒ Mean,

要求计算均数、标准差和中位数

☐ Continue

分析中用到的操作界面如图 11.15 所示。

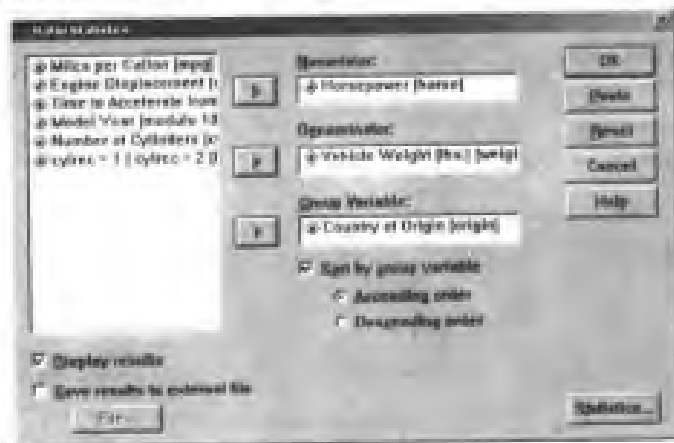


图 11.15 主对话框

【主对话框】（见图 11.15）

1. Numerator 框：用于选入作为相对比分子的变量。
2. Denominator 框：用于选入作为相对比分母的变量。
3. Group Variable 框：选入分组变量，相对比指标将分组进行计算。
4. Sort by group variable 框：选入分组变量时可用，要求将数据按照分组变量排序。

5. Display results: 要求在结果窗口中输出分析结果, 系统默认。
6. Save results to external file: 要求将分析结果存为外部数据文件。

【Statistics 子对话框】(见图 11.16)

该对话框中提供了许多比较专业的相对比描述指标。

1. Central Tendency 复选框组: 选择用于描述相对比集中趋势的指标, 有中位数、均数和加权均数三种, 其中加权均数的算法为分子的均数除以分母的均数(意思等同于以分母大小为权重)。下方还可选择输出相应指标设定范围的可信区间。

2. Dispersion 复选框组: 选择用于描述相对比离散趋势的指标, 除了大家都熟悉的标准差、全距、最小值、最大值以外, 还有几个专用指标如下。

- ✧ AAD: 即平均绝对离差(Average Absolute Deviation), 全部相对比与相对比中位数差值的绝对值之和除以样本数。
- ✧ COD: 离散系数(Coefficient of Dispersion), 相对比的平均绝对离差与相对比中位数的比值。
- ✧ PRD: 价格相关微分(Price-related Differential), 均数除以加权均数。
- ✧ Median centered COV: 基于中位数的变异系数, 计算方法为各相对比与中位数差值的均方根除以中位数。
- ✧ Mean centered COV: 基于均数的变异系数, 即通常所使用的变异系数, 计算方法为标准差除以均数。

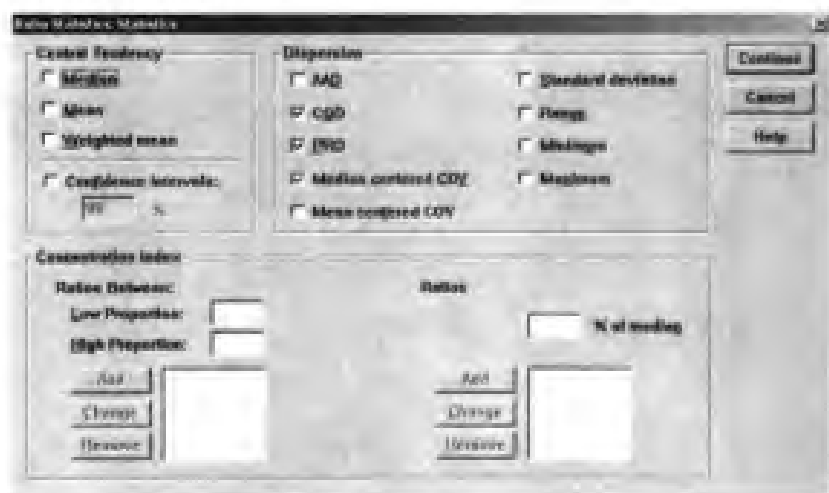


图 11.16 Statistics 子对话框

3. Concentration Index 框组: 用于计算集中系数(Coefficient of Concentration), 该指标用于描述相对比大小落入某一区间所占的比例, 它有两种计算方式。

- ✧ Ratios Between: 由用户自行定义具体的区间上、下界, 显然界值需要成对输入, 可以定义多对界值, 系统将分别计算相应比例。
- ✧ Ratios Within: 由用户指定上下界距离中位数的百分比, 需要键入 0-100 的数值, 相应的区间下界等于 $(1-0.01 \times \text{value}) \times \text{中位数}$, 上界等于 $(1+0.01 \times \text{value}) \times \text{中位数}$ 。

11.4.2 结果解释

Ratio Statistics

Case Processing Summary

		Count	Percent
Country of Origin	American	249	62.4%
	European	71	17.8%
	Japanese	79	19.8%
Overall		399	100.0%
Excluded		7	
Total		406	

首先给出的是纳入分析的记录报告，可见在总共 406 条纪录中，有 7 条因相关变量存在缺失值而未能纳入分析，最终用于分析的记录为 399 条。

Ratio Statistics for Horsepower / Vehicle Weight (lbs.)

Group	Mean	Price Related Differential	Coefficient of Dispersion	Coefficient of Variation
				Median Centered
American	.035	.991	.136	18.5%
European	.033	1.005	.140	18.6%
Japanese	.036	.993	.102	12.8%
Overall	.035	.992	.130	17.4%

上表给出的就是不同产地的汽车功率和车重之比，可见美国、欧洲和日本车的相对比均数基本接近，在离散度指标上，PRD 相差不大，但日本车的 COD 明显较小，中位数变异系数的大小情况也与之相似，可见日本车相对比的离散度较低。

11.4.3 对引例的进一步分析

如果不使用 Ratio 过程，而是先计算出一个新变量代表每辆车的比值，然后对该变量进行描述，则可以得到如下分析结果：

Descriptive Statistics

Country of Origin		N	Minimum	Maximum	Mean	Std. Deviation
	TEMPVAR	1	.13	.13	.1270	
	Valid N (listwise)	1				
American	TEMPVAR	249	.02	.07	.0351	.00625
	Valid N (listwise)	249				
European	TEMPVAR	71	.02	.05	.0333	.00632
	Valid N (listwise)	71				
Japanese	TEMPVAR	79	.03	.05	.0357	.00424
	Valid N (listwise)	79				

显然，分析结果是完全相同的，但是使用这种方法只能计算出常用描述指标，而 Ratio 过程可以提供许多专门的相对比描述指标，使用上要方便的多。

11.5 综合分析实例

11.5.1 频数表数据

例 11.3 某市 1995 年 110 名 7 岁男童的身高资料已按频数表格式输入 (数据 high.sav), 变量 groupmid 代表所在组段的组中值, freq 代表组段频数, 请求出该资料的均数、标准差、中位数和四分位数间距。

解: 首先, 数据集中的每一条记录代表的不是一个观察对象, 而是一个频数组, 因此要让 SPSS 正确识别该数据, 就应当先用 Weight Cases 过程将变量 Freq 值定为频数变量, 然后才能进行统计; 其次, 虽然所有这四项指标 Frequencies 过程和 Explore 过程都可以求得, 但计算百分位数时要求将数据按频数表的形式来分析, 否则就会出错, 这只能在 Frequencies 过程中用 Values are group midpoints 复选框来实现。因此, 我们的操作步骤如下:

Data → Weight Cases

☒ Weight cases by

Frequency variables 框: Freq

OK

Analyze → Descriptive Statistics → Frequencies

Variables 框: groupmid

Statistics:

☒ Quartiles, ☒ Std. deviation, ☒ Mean

☒ Values are group midpoints

Continue

OK

要求设定频数变量

指定频数变量为 freq

指定要描述的变量为 groupmid

要求计算四分位数、均数和标准差

提醒系统数据是组中值

系统会给出相应的分析结果如左下表, 可见其中三个百分位数的结果是正确的。右下表是不选中 Values are group midpoints 复选框的结果, 可见三个百分数的计算结果不对, 但均数和标准差则是正确的。

Statistics			Statistics		
身高组中值			身高组中值		
N	Valid	110	N	Valid	110
	Missing	0		Missing	0
Mean		121.9455	Mean		121.9455
Std. Deviation		4.7213	Std. Deviation		4.7213
Percentiles	25	118.6667 ^a	Percentiles	25	119.0000
	50	121.9231		50	122.0000
	75	124.9429		75	125.0000

a. Percentiles are calculated from grouped data.

11.5.2 偏态分布数据的参考值范围

例 11.4 请求出数据 brain.sav 中的变量 age 的 95% 参考值范围。

解: 求参考值范围有两种方法, 当数据服从正态分布时可从正态分布原理求出,

否则就应当用百分位数法求出。该题我们如果先对变量 age 作直方图，就会发现它的分布明显呈正偏态，因此只能用后者。所以，这个问题最终归结到计算 P2.5 和 P97.5 两个百分位数上了，显然，这对 Frequencies 过程来说易如反掌。操作如下：

Analyze→Descriptive Statistics→Frequencies

Variables 框: age

欲分析的变量为 age

Statistics:

☒ Percentiles: 键入 2.5; 键入 97.5;

计算 P2.5 和 P97.5 百分位数

Charts:

☒ Histograms

做出直方图，验证分布的偏态

最终的统计分析结果如左下所示，可知 95% 参考值范围是 17~77.25 岁。而图 11.17 的直方图则显示出年龄的确不服从正态分布。

Statistics		
AGE		
N	Valid	189
	Missing	0
Percentiles	2.5	17.00
	97.5	77.25

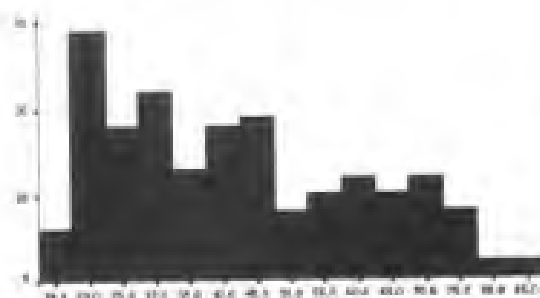


图 11.17 输出的直方图

11.5.3 标准正态变换

例 11.5 请将数据 xuelin.sav 中的变量 X 分正常人组和克山病人组做标准正态变换。

解：标准正态变换是多元分析中非常常用的手段，实现它的方法也很多，最基本的就是求出均数和标准差后用 Compute 过程计算，但 Descriptives 过程已经提供了将数据的标准正态变换结果存为新变量的功能，我们直接利用就可以了。但是，本题要求按变量 group 的取值分组计算，Descriptives 过程并不直接支持分组，因此首先我们要使用 Split file 过程将数据一分为二后才能继续分析。操作如下：

Data→Split file

☒ Compare groups

Groups based on 框: group

要求分析时按 group 的取值分组分析

Analyze→Descriptive Statistics→Descriptives

Variables 框: X

指定描述的变量为 x

☒ Save standardized values as variables

要求存储标准正态变换结果

本题的输出结果除了在结果窗口中的输出外，还会在数据集中生成一个新的变量，

该变量自动命名为“Z+原变量名”，此处即为 ZX。就是 X 标准正态变换后的数值，如下表所示。表中的 zsc001 是按不分组情况作的标准正态变换，可见两者是不同的。

X	GROUP	ZX	ZSC001	X	GROUP	ZX	ZSC001
.84	1.00	-1.61432	-.94862	.54	2.00	-1.29010	-1.58875
1.05	1.00	-1.11645	-.50054	.64	2.00	-1.05322	-1.37537
1.20	1.00	-.76082	-.18048	.64	2.00	-1.05322	-1.37537
1.20	1.00	-.76082	-.18048	.75	2.00	-.79265	-1.14066
1.39	1.00	-.31036	.22493	.76	2.00	-.76896	-1.11932
1.53	1.00	.02155	.52365	.81	2.00	-.65052	-1.01264
1.67	1.00	.35347	.82238	1.16	2.00	.17857	-.26583
1.80	1.00	.66168	1.09976	1.20	2.00	.27333	-.18048
1.87	1.00	.82764	1.24913	1.34	2.00	.60496	.11824
2.07	1.00	1.30180	1.67587	1.35	2.00	.62865	.13958
2.11	1.00	1.39664	1.76122	1.48	2.00	.93660	.41697
				1.56	2.00	1.12610	.58767
				1.87	2.00	1.86044	1.24913

11.5.4 探索性分析

请注意，本例比较复杂，里面使用到了许多综合性的知识，请大家一定认真阅读，完全理解本例将会对大大提高朋友们的统计分析能力。

例 11.6 请分析 SPSS 自带数据文件 Anxity.sav 中评分 (Score) 的分布情况如何，以及四次实验 (trial) 间的评分有无变化趋势、方差是否齐。

解：一般来说，评分的分布情况是不太明确的，一般都不认为它服从正态分布（想想自己班上的期末考试成绩分布吧），因此在这里我们要做的是个典型的探索性分析，具体来说，我们主要关心：1. 该评分是否服从正态（或对称）分布；2. 有无异常值出现。解决这些问题用 Explorer 过程是非常合适的，可以用它做出常用指标以及箱式图、直方图来判断。当然，这一目的可以用许多方法达到，这里只举出用 Explorer 过程的做法。

再来看第二个目的：研究四次 trial 间的评分有无变化趋势，该目的同样可以再调用一次 Explorer 过程达到，只需要将变量 trial 作为一个 factor 纳入，所作的箱式图就会直观的给出变化趋势来。

最后是第三个分析目的：四组间方差是否齐，这可以直接使用 Spread vs. Level with Levene Test 单选框组来实现，首先我们可以使用 Untransformed 单选框，将使用原始数据来分析，再根据结果来决定是否进行变换。

最终的操作如下：

Analyze→Descriptive Statistics→Explore

Dependent variables 框：Score

Statistics:

☒ Outliers

要求列出离群值

☒ Continue

Plots:

☒ Histogram

做出直方图

☒ Normality with plot test

要求做出正态分布图，进行正态性检验

Continue

OK

Analyze→Descriptive Statistics→Explore

Dependent variables 框: Score

Factor 框: Trial

按 Trail 的不同取值分组分析

Plots: Spread vs. Level with Levene Test:

Untransformed

要求使用原始数据进行方差齐性分析

Continue

OK

主要的输出结果如下所示:

Explore

Descriptives					Extreme Values				
Score	Mean	Statistic	Std. Error		Score	Highest	Case Number	Value	
	Mean	10.00	.75			1	5	18	
	95% Confidence Interval for Mean	8.50				2	37	19	
		Upper Bound				3	1	18	
		11.50				4	29	18	
	5% Trimmed Mean	10.00				5	21	18	
	Median	10.00				Lowest	1	32	1
	Variance	26.706				2	24	1	
	Std. Deviation	5.17				3	12	2	
	Minimum	1				4	20	2	
	Maximum	19				5	36	2	
	Range	18							
	Interquartile Range	8.00							
	Skewness	.038	.343						
	Kurtosis	-.951	.674						

上面的两个表格分别为描述表格和极端值统计表。从左上方的描述表格可见:均数为 10, 和中位数完全相同, 偏度系数也接近于 0, 这两点共同表明数据分布的对称性非常好; 最小、最大值分别为 1 和 19, 离均数 10 的距离相似, 去除 5% 极端值后的均数也为 10, 共同说明数据中不大可能有离群值; 峰度系数的值虽然较大, 但其标准误差也较大, 提示研究该数据分布时唯一要注意的就是其峰度。右侧的极端值统计表列出了最大和最小的各 5 个值, 显然不存在离群值。

Tests of Normality

Score	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Score	.106	48	.200*	.946	48	.047

* This is a lower bound of the true significance.

^a Lilliefors Significance Correction

上表分别做了两个正态性检验: Kolmogorov-Smirnov 检验不拒绝 H_0 , Shapiro-Wilk 检验却刚好拒绝了 H_0 , 这可让我们如何是好? 别急, 继续往下看。

图 11.18 是做出的直方图和箱式图, 直方图可见对称性尚可, 但峰度是稍差了一些。箱式图因为主要显示对称性, 所以效果非常的好。

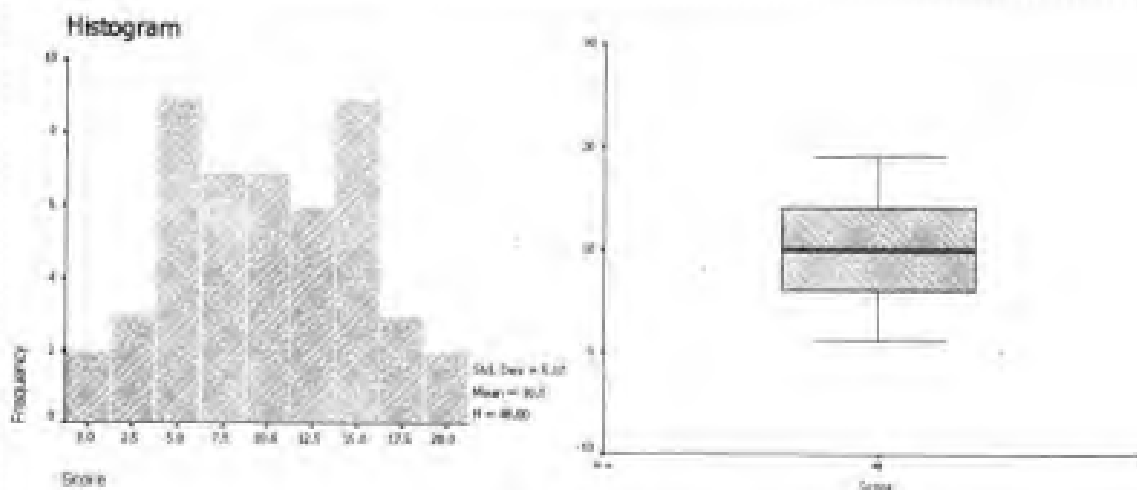


图 11.18 直方图和箱式图

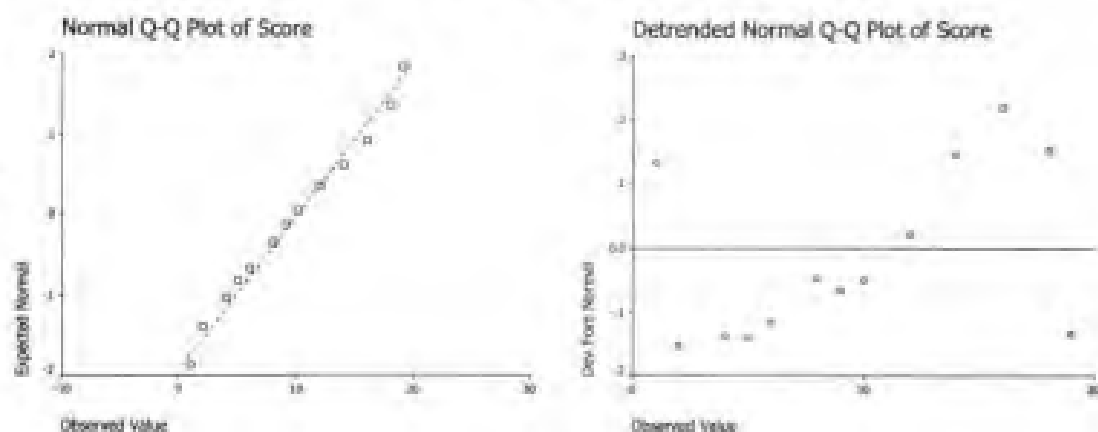


图 11.19 QQ 图和去势 QQ 图

图 11.19 是用于正态检验的 QQ 图和去势 QQ 图（QQ 残差图）。从 QQ 图可见数据基本上还是呈正态走势，只是有非常轻微的波动，QQ 残差图将这种波动放大了，可见这种波动基本上都在两个标准差范围内，是可以接受的。因此，可以认为该数据服从正态分布。

下面是分四次实验的输出，由于内容太多，就不一一列出了，只给出关键部分如下：

Test of Homogeneity of Variance					
		Levene Statistic	df1	df2	Sig.
Score	Based on Mean	.614	3	44	.610
	Based on Median	.492	3	44	.690
	Based on Median and with adjusted df	.492	3	43.094	.690
	Based on trimmed mean	.537	3	44	.680

上表为稳健 Levene 方差齐性检验的结果，分别给出了基于均数、中位数、中位数并调整自由度、删除极端值后均数的检验结果。可见在这几种情况下 P 值均在 0.6 以

上，因此数据的方差齐性情况比较好，可以进行下一步的分析。

图 11.20 为分组箱式图，从中可见随着试验序号的增加，评分呈逐渐下降趋势，即越往后，评分越低，且四组数据的变异程度比较接近。

图 11.21 为分布——水平图，可见随着中位数的上升，四分位数间均呈现下降趋势，相应的斜率为-0.275。但由于前面方差齐性检验已经得出了方差齐的结论，此处我们不再需要对此问题进行进一步的分析了。

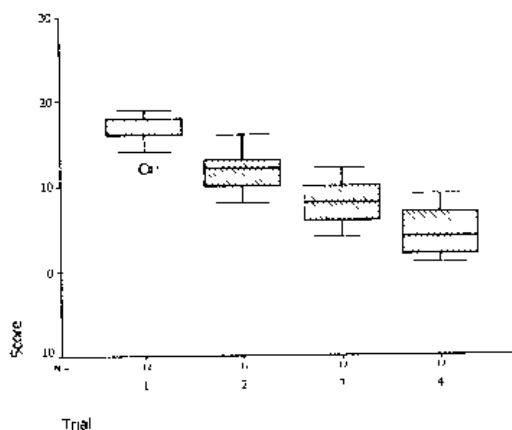


图 11.20 分组箱式图

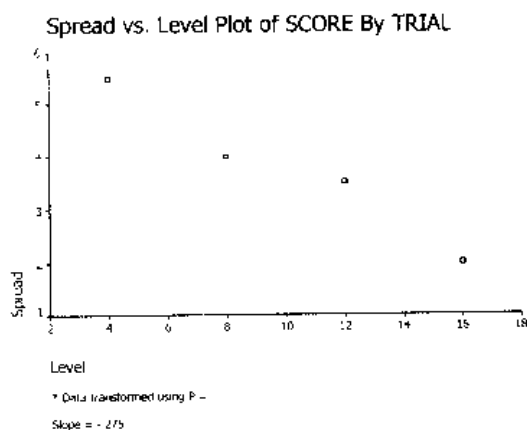


图 11.21 分布——水平图



朋友们如果有兴趣，可以进入图形编辑模式，在图 11.20 的纵轴上添加参考线，立刻就会发现图 11.21 上的中位数和四分位数间距可以和图 11.20 中一一对应起来，这可以帮助大家理解分布——水平图的含义。

第 12 章 均数间的比较

——Compare Means 菜单详解

无论数据有多复杂，许多临床研究生的毕业论文中统计方法无外乎 t 检验和卡方检验两种，这一方面说明我们普及正确统计方法的任务还非常繁重，但也从另一个方面证明了这两种方法的重要地位。


——张文彤

现在计算机技术的发展速度可真快呀，就拿 CPU 来说吧，前些天还在“奔散”呢，现在又要“奔死”了，说不定一年后就该“奔无”了吧。可是你知道吗？在奔腾及更早的 CPU 所采用的 CISC 指令集中有个著名的 80/20 规则，也就是有 80% 的任务是被 20% 的最常用指令所完成的；换言之，另外 80% 的复杂指令只完成 20% 的不常用任务。

好了，言归正传。现在我要非常高兴的向大家宣布：80/20 规则在 SPSS 的使用中同样有效！仅以 Analyze 菜单为例，其中最常用的子菜单为：

- ◇ Descriptive Statistics
- ◇ Compare Means
- ◇ General Linear Model (第一项)
- ◇ Correlate (第一项)
- ◇ Regression (前半截)

只要掌握了它们的使用秘籍，你就可以理直气壮的宣称你已经学会了如何用 SPSS 解决 80% 的统计学问题。如果不满足，在召开新闻发布会的时候你还可以对以上指标进行四舍五入。

 此时课堂上有一漂亮 MM 提问：老师，那我们是不是只学这几项功能就行了？我...我...气死我了...

好，言归正传。在以上五个菜单中，Compare Means 是最简单的一个，但使用频率却几乎最高！因此，它的重要性也就不用我多说了吧。

现在就让我们大家一起踏上 Compare Means 之旅。该菜单集中了几个用于计量资料均数间比较的过程。具体有：

- ◇ Means 过程：该过程实际上更倾向于对样本进行描述，它可以对需要比较的各组计算描述指标，进行检验前的预分析。当然如果你愿意，也可直接比较。
- ◇ One-Samples T Test 过程：进行样本均数与已知总体均数的比较。
- ◇ Independent-Samples T Test 过程：进行两样本均数差别的比较，即通常所说的两组资料的 t 检验。
- ◇ Paired-Samples T Test 过程：进行配对资料的均数比较，即配对 t 检验。
- ◇ One-Way ANOVA 过程：进行两组及多组样本均数的比较，即成组设计的方差

分析，还可进行随后的两两比较。

细心的朋友可能会发现，这里好像没有 u 检验什么事儿。因为 u 检验是在大样本情况下使用，而此时 t 检验的结果已经和它基本一致，所以 SPSS 没有在菜单上专门为它留出位置来。

12.1 Means 过程

和上一章所讲述的几个专门的描述过程相比，Means 过程的优势在于所有的描述统计量均按自变量的取值分组计算，无需像其他过程那样必须先调用 Split File 过程。在输出结果中各组的描述指标被放在一起，也便于相互比较。如果需要，Means 过程还可以直接输出方差分析结果，并计算相应的相关性指标而无须再次调用其他过程，显然要比先描述再比较方便的多。

12.1.1 引例

例 12.1 试初步分析数据文件 pkc.sav 中肿瘤病人的性别及分期对 PKC 的值有无影响。

解：由于是初步分析，我们需要分性别、分期对 PKC 做出描述，并在可能的情况下给出检验结果，这用 Means 过程只需要一次就完成了。

Statistics→Compare Means→Means

Dependent List 框: pkc

指定需要分析的变量为 pkc

Independent List 框: sex、jibie

分组变量为 sex、jibie

Options

☒ Anova table and eta

要求作方差分析，并计算 eta 值

Continue

OK

12.1.2 界面说明

【主对话框】（见图 12.1）

1. Dependent List 框：用于选入需要分析的变量，如果选入两个以上变量，系统会在同一张输出表各种依次给出其分析结果。本题选入变量 PKC。

2. Layer 按钮组：和 Crosstabs 过程中遇到的 Layer 按钮组类似，共包括 Previous 和 Next 两个按钮，用于控制下方的 Independent List 框。如果将自变量分入了不止一层，则最后系统在分析时会按各个自变量的组合情况来给出结果，如男*二期，女*二期；否则会依次分析它们对结果变量的影响情况。本例要求对性别及分期进行初步分析，因此最好是单独分析它们和变量 PKC 的关系，即不分层。

3. Independent List 框：用于选入分组变量，如果选入两个以上变量，系统会根据 Layer 的设置情况做出不同反应，见上。本题应选入变量 sex 和 jibie。

【Options 子对话框】（见图 12.2）

用于选择需要计算的描述统计量和统计分析，包括：

1. Statistics 框：可选的描述统计量。它们是：

- ◇ Sum, Number of Cases: 总和, (进入分析的) 记录数。
- ◇ Mean, Geometric Mean, Harmonic Mean: 均数, 几何均数, 调和均数。
- ◇ Standard Deviation, Variance, Standard Error of The Mean: 标准差, 方差, 标准误。
- ◇ Median, Grouped Median: 中位数, 频数表资料中位数 (比如 30 岁组有 5 人, 40 岁组有 6 人, 则在计算 Grouped Median 时均按组中值 35 和 45 进行计算)。
- ◇ Minimum, Maximum, Range: 最小值, 最大值, 全距。
- ◇ Kurtosis, Standard Error of Kurtosis: 峰度系数, 峰度系数的标准误。
- ◇ Skewness, Standard Error of Skewness: 偏度系数, 偏度系数的标准误。
- ◇ Percentage of Total Sum, Percentage of Total N: 总和的百分比, 占总样本例数的百分比。

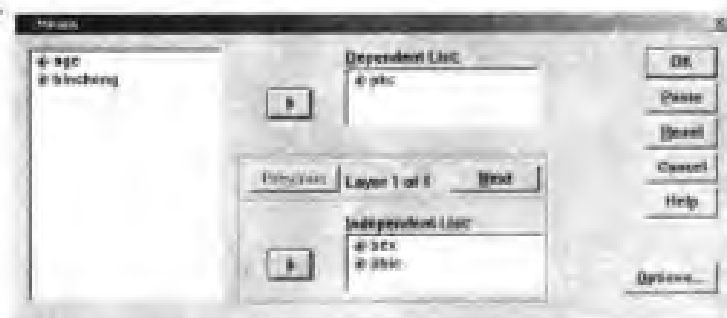


图 12.1 主对话框



图 12.2 Options 子对话框

2. Cell Statistics 框: 选入的描述统计量, 默认情况下系统已经选入了均数、样本例数和标准差三种。

3. Statistics for First layer 复选框组: 用于选择是否检验第一层 (注意是第一层而不是第一个) 的分组变量对结果变量的影响有无统计学意义。

- ◇ Anova table and eta: 对分组变量进行单因素方差分析, 并计算 eta 值。它用于度量分组变量和结果变量间的关联性, eta 的平方表示由组间差异所解释的结果变量的方差的比例, 即 SS 组间/SS 总。
- ◇ Test for linearity: 检验线性相关性, 即不同组的均数间是否存在线性趋势。实际上也就是进行单因素方差分析, 但同时会计算出 R 和 R^2 , 前者是大家熟悉的相关系数, 后者即决定系数。请注意, 如果第一层中有多个自变量, 则 SPSS 只对最后一个自变量计算 R 和 R^2 。

12.1.3 结果解释

有了上一章的基础, Means 过程的输出看起来就不太困难了。本例输出如下:

Means

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
PKC * SEX	51	100.0%	0	.0%	51	100.0%
PKC * JIBIE	51	100.0%	0	0%	51	100.0%

上表还是缺失值报告，不过由于是两个自变量，共分了两行来给出结果。

PKC * SEX

Report

PKC			
SEX	Mean	N	Std. Deviation
1	66.5647	34	15.2316
2	64.1941	17	16.4050
Total	65.7745	51	15.5085

首先给出的是最先选入的自变量 sex 的分析结果，上表为常用统计描述量报表。这里按默认情况输出均数，样本量和标准差。由于我们选择了分组变量，因此三项指标均给出分不同性别的分组及合计值，从中可见女性的 PKC 均数稍低。显然，以这种方式列出统计量可以非常直观的进行各组间的比较。

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
PKC * SEX	Between Groups	(Combined)	63.690	1	63.690	.261	.612
	Within Groups		11962.007	49	244.123		
	Total		12025.697	50			

上表为单因素方差分析表。在选择了 Anova table and eta 或 Test for linearity 复选框时出现。实际上就是在检验各组间均数有无差异，可见性别对 PKC 无影响。上面各项的具体含义将在单因素方差分析一节中解释。

Measures of Association

	Eta	Eta Squared
PKC * SEX	.073	.005

相关性度量指标，给出 Eta 值以及 Eta 值的平方根。可见性别和 PKC 指标的相关度非常差，和前面的分析结果相一致。

PKC * JIBIE**Report****PKC**

JIBIE	Mean	N	Std. Deviation
2.00	80.2350	20	6.4556
3.00	65.1533	15	7.9035
4.00	48.2812	16	9.2204
Total	65.7745	51	15.5085

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
PKC * JIBIE	Between Groups	(Combined)	9084.130	2	4542.065	74.117	.000
	Within Groups		2941.567	48	61.283		
	Total		12025.697	50			

Measures of Association

	Eta	Eta Squared
PKC * JIBIE	.869	.755

上面的一系列结果为分期和 PKC 指标的关系, 可见不同分期的 PKC 值是不同的, 从样本均数上看趋势为分期越高, PKC 越低。

12.1.4 对引例的进一步分析**【线性趋势的检验】**

如果在 **Options** 子对话框中选择了 **Test for linearity** 复选框, 则 PKC*JIBIE 的输出中会得到如下表格:

Measures of Association

	R	R Squared	Eta	Eta Squared
PKC * JIBIE	-.869	.755	.869	.755

依次给出了相关系数、决定系数、eta 值和 eta 值的平方, 可见不同 jiebie 组的 pkc 取值间的线性趋势还是比较明显的。但这只是初步分析, 具体的情况还要通过均数图以及进一步的分析来验证。

PKC*SEX 的分析结果不会有任何变化, 即 SPSS 在这里只分析第一层中最后一个分组变量的线性趋势。

【分层选入分组变量】

如果选入分组变量的时候不是选入同一层, 而是分别选入两层, 则最后分析结果中的 Report 表如下:

Report

PKC				
SEX	JIBIE	Mean	N	Std. Deviation
1	2.00	80.4462	13	6.4875
	3.00	66.7000	10	7.8909
	4.00	50.0364	11	10.3938
	Total	66.5647	34	15.2316
2	2.00	79.8429	7	6.8934
	3.00	62.0600	5	7.7838
	4.00	44.4200	5	4.6499
	Total	64.1941	17	16.4050
Total	2.00	80.2350	20	6.4556
	3.00	65.1533	15	7.9035
	4.00	48.2812	16	9.2204
	Total	65.7745	51	15.5085

即按自变量的各种组合分别给出描述结果。但后面的检验结果不变。


12.2 One-Samples T Test 过程

One-Samples T Test 过程用于进行样本所在总体均数与已知总体均数的比较，即单样本的 t 检验。由于样本数据是通过随机调查若干名观察对象得来，我们只知道它所在总体的均数在该样本均数的附近，但具体是多少并不清楚。为了回答该问题，统计学上采用了小概率反证法的原理：我们有如下两种假设：

$H_0: \mu = \mu_0$ ，样本均数与总体均数的差异完全是抽样误差造成。

$H_1: \mu \neq \mu_0$ ，样本均数与总体均数的差异除了由抽样误差造成外，也反映了两个总体均数确实存在的差异。

显然两者中必然会有一个是对的，究竟是哪一个是呢？我们不妨先假设是 H_0 成立，即一切的一切都是抽样误差惹的祸。在这个前提下，我们的样本是从已知均数的大总体中抽出来的，那么从这个总体中抽出这样一个样本均数（以及更极端情况）的概率为多少呢？这可以通过统计学方法计算出来，即我们所求得的 P 值。如果该 P 值太小，成为了我们所定义的小概率事件（小于等于 α 水准），则我们怀疑所做的假设不成立，从而拒绝 H_0 ，投向 H_1 的怀抱；反之，我们就不能拒绝 H_0 ，但一般也不太好说去接受它。

 打个比方吧，拒绝 H_0 相当于证据确凿，判处死刑。不拒绝 H_0 则相当于证据不足，当庭释放。听清楚了，法庭可没说你是彻底清白的，是因为证据不足才释放，哪天凑够了证据还要来逮你的！

具体计算 P 值的检验方法有很多, 当样本所在总体服从正态分布时, 就可以使用这里提到的 t 检验。

12.2.1 引例

例 12.2 根据大量调查, 已知某地成年男子脉搏均数为 72 次/分, 现在该地邻近的山区随机调查了 20 名健康成年男子, 测得其脉搏值如下, 请据此推断山区成年男子的脉搏均数是否与该地成年男子有所不同, 数据见 pulse.sav。

测量值: 75 74 72 74 79 78 76 69 77 76 70 73 76 71 78 77 76 74 79 77

解: 20 名男子显然是从山区总体中抽得的一个样本, 现在要比较的就是他所在的山区总体均数是否等于已知的 72 次/分。显然, 用 One-Samples T Test 过程再合适不过了, 只需要自行定义已知总体均数即可。

Statistics→Compare Means→One-Samples T Test

Test Variable(s)框: pulse

分析的结果变量为 pulse

Test Value:框: 键入 72

已知总体均数为 72 次/分

OK

12.2.2 界面说明

【主对话框】(见图 12.3)

1. Test Variables 框: 用于选入需要分析的变量。
2. Test Value 框: 用于输入已知的总体均数, 默认值为 0。

【Options】子对话框(见图 12.4)

用于定义相关的选项。

1. Confidence Interval 框: 输入需要计算的均数差值可信区间范围, 默认为 95%, 可自行更改。如果是和总体均数为 0 相比, 则此处计算的就是样本所在总体均数的可信区间。

2. Missing Values 单选框组: 定义分析中对缺失值的处理方法, 可以是具体分析用到的变量有缺失值才去除该记录(Excludes cases analysis by analysis), 或只要相关变量有缺失值, 则在所有分析中均将该记录去除(Excludes cases listwise)。默认为前者, 以充分利用数据。



图 12.3 主对话框



图 12.4 Options 子对话框

12.2.3 结果解释

One-Samples T Test 过程的输出也是比较简单的，由描述统计表和 t 检验表组成，上题输出结果如下：

T-Test

One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
PULSE	20	75.050	2.8924	.6468

所分析变量的基本情况描述，有样本量、均数、标准差和标准误。注意在 10.0 版的分析结果中均数和标准差的小数要比上表少一位。


One-Sample Test						
Test Value = 72						
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
PULSE	4.716	19	.000	3.050	1.696	4.404

上表为单样本 t 检验表，第一行注明了用于比较的已知总体均数为 72，下面从左到右依次为 t 值(t)、自由度(df)、P 值(Sig.2-tailed)、两均数的差值(Mean Difference)、差值的 95%可信区间。由上表可知： $t=4.716$ ， $P<0.001$ 。因此可以拒绝 H_0 ，接受 H_1 ，认为山区健康成年男子的脉搏均数与该地成年男子有所不同，结合具体的均数值，可以认为山区较高。

12.3 Independent-Samples T Test 过程

Independent-Samples T Test 过程用于进行两样本均数的比较，即常用的两样本 t 检验。和上面单样本 t 检验的原理相同，我们也采用了小概率反证法，首先假设 H_0 ：两样本来自同一总体。当该总体服从正态分布时，我们就可以采用两样本 t 检验来计算从该总体中抽得这样两个活宝的概率为多少，从而做出统计推断。

由于 H_0 假设的是两样本来自同一总体，因此两样本 t 检验在推导过程中除了要求总体服从正态分布外，还要求两样本各自所在总体方差相同（不然怎么可能是同一总体呢）。如这些应用条件不被满足，情况较轻时可以采用校正 t 检验的结果，否则应使用变量变换使之满足条件，或采用非参数检验过程。

 t 检验对数据稍微偏离应用条件有较好的耐受性，所以分析时往往无需严格检验分布情况，肉眼估计即可。

12.3.1 引例

例 12.3 某医生测得 18 例慢性支气管炎患者及 16 例健康人的尿 17 酮类固醇排出量(mg/dl)分别为 X_1 和 X_2 ，试问两组的均数有无不同（倪宗璜，《医学统计学》第二

版 P19)。

X1: 3.14 5.83 7.35 4.62 4.05 5.08 4.98 4.22 4.35 2.35 2.89 2.16 5.55 5.94 4.40 5.35 3.80 4.12

X2: 4.12 7.89 3.24 6.36 3.48 6.74 4.67 7.38 4.95 4.08 5.34 4.27 6.54 4.62 5.92 5.18

解：该数据为定量资料，设计为成组设计，目的是两样本均数的比较，显然用两样本均数比较的 t 检验是非常合适的，正态性可以先作直方图大致看一下（不分组），方差齐性检验则会在 t 检验结果中自动给出。根据分析要求，数据输入时应有两个变量，见数据文件 guchun.sav，变量 guchun 为测得的类固醇排出量，group 则用于区分该观察对象为病人还是健康人。

Graphs→Histogram

Variable 框: guchun

要求对变量 guchun 作直方图

OK

Analyze→Compare Means→Independent-Samples T Test

Test Variable(s)框: guchun

要分析的变量为 guchun

Grouping Variable 框: group

分组变量为 group

选中变量 group: Define Groups:

定义检验的两组

Group1: 键入 1

Group=1 的一组进入检验

Group2: 键入 2

Group=2 的一组进入检验

Continue

OK

12.3.2 界面说明

【主对话框】（见图 12.5）

1. Test Variables 框：用于选入需要分析的变量。
2. Grouping Variable 框：用于选入分组变量。注意选入后还要定义需比较的组别。


【Define Groups 子对话框】（见图 12.6）

用于定义需要相互比较的两组的分组变量值。

1.  Use Specified values: 指定分组变量的两个取值，相应的两组将进行比较。



可以这样来理解：如果分组变量有 3 个取值（即有三组），而我们做 t 检验是比较其中的某两组，这时就可以用 Use Specified values 来指定需比较的两组。当然，如果分组变量只有 2 个取值时，我们仍然要在该框中进行定义，这也算是 SPSS 对话框存在的一个小缺陷吧。

2.  Cut point: 如果没有明确的分组变量，而是按照某个取值的分界线来进行比较（如小于 30 岁的和大于等于 30 岁的进行比较），就可以用 Cut Point 来指定。右侧的数值框用于输入分界值，系统会将记录自动分成小于界值和大于等于界值的两组来进行比较。

【Options 子对话框】

和 One-Samples T Test 对话框的 Options 完全相同，此处不再重复。

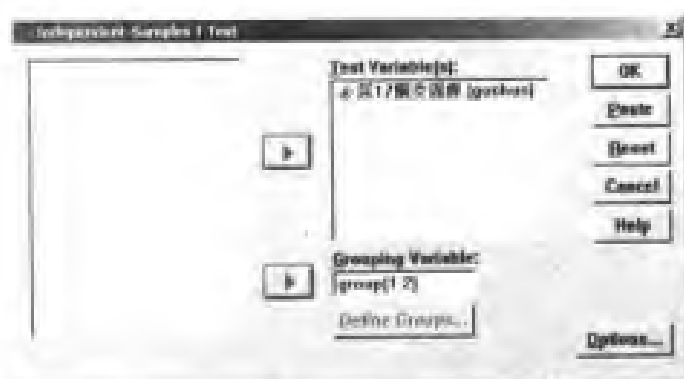


图 12.5 主对话框



图 12.6 Define Groups 子对话框

12.3.3 结果解释

上题中 t 检验的结果输出如下:

T-Test

Group Statistics				
GROUP	N	Mean	Std. Deviation	Std. Error Mean
尿17酮类固醇 慢性炎患者	18	4.4544	1.3245	.3122
健康人	16	5.2988	1.3820	.3455

两组需检验变量的基本情况描述。


Independent Samples Test									
	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
X Equal variances assumed	.225	.638	-1.818	32	.078	-.8443	.4644	-1.7904	.1017
Equal variances not assumed			-1.813	31.163	.079	-.8443	.4656	-1.7938	.1052

该结果分为两大部分: 第一部分为 Levene's 方差齐性检验, 用于判断两总体方差是否齐, 这里的检验结果为 $F = 0.225$, $P = 0.638$, 可见在本例中方差是齐的; 第二部分则分别给出两组所在总体方差齐和方差不齐时的 t 检验结果, 由于前面的方差齐性检验结果为方差齐, 第二部分就应选用方差齐时的 t 检验结果, 即上面一行列出的 $t = -1.818$, $v = 32$, $P = 0.078$, 从而最终的统计结论为按 $\alpha = 0.05$ 水准, 不拒绝 H_0 , 尚不能认为慢性支气管炎患者的尿 17 酮类固醇排出量与健康人不同。最后面还附有一些其他指标, 如两组均数差值的可信区间等, 以对差异情况有更直观的了解。

上表的标题内容翻译如下:

	Levene 方差齐性检验				两均数是否相等的 t 检验				
	F 值	P 值	t 值	自由度	P 值 (双侧)	均数差值	差值的标准误	差值的 95%可信区间	
X	假设方差齐	.225	.638	-1.818	32	.078	-.8443	.4644	-1.7904 .1017
	假设方差不齐				-1.813	31.163	.079	-.8443	.4656 -1.7938

 如果你觉得上表太宽，用第 3 章学过的行列转置功能可以使它变的紧凑许多。

 虽然上表也给出了方差不齐时 t 检验的修正结果，但当 t 检验应用条件被严重违反时，该结果也是不准确的，现在统计学界比较一致的看法是此时应当直接采用非参数检验，这样虽然损失一些检验效能，但不会出大错。

12.4 Paired-Samples T Test 过程

该过程用于进行配对设计的差值均数与总体均数 0 比较的 t 检验，配对设计有两种情况：1. 对同一个受试对象处理前后的比较，这种设计由于在结果中混杂了时间因素的影响，现在已不推荐使用；2. 将受试对象按情况相近者配对（或者自身进行配对），分别给予两种处理，以观察两种处理效果有无差别。在配对设计得到的样本数据中，每对数据之间都有一定的相关，如果采用成组的 t 检验就无法利用这种关系，浪费了大量统计信息。

对于这种情况，统计学上的解决办法是求出每对的差值：如果两种处理实际上没有差异，则差值的总体均数应当为 0，从该总体中抽出的样本其均数也应当在 0 附近波动；反之，如果两种处理有差异，差值的总体均数就应当远离 0，其样本均数也应当远离 0。这样，通过检验该差值总体均数是否为 0，就可以得知两种处理有无差异。

对统计学比较熟悉的朋友可以看出，Paired-Samples T Test 过程的功能实际上是和 One-Samples T Test 过程相重复的（等价于已知总体均数为 0 的情况），但 Paired-Samples T Test 过程使用的数据输入格式和前者不同，因此它仍然有存在的价值。

12.4.1 引例

例 12.4 为研究女性服用某避孕新药后是否影响其血清总胆固醇，将 20 名女性按年龄配成 10 对。每对中随机抽取一人服用新药，另一人服用安慰剂。经过一定时间后，测得血清总胆固醇含量(mmol/L)，结果如下表。问该新药是否影响女性血清总胆固醇（倪宗瓚，《卫生统计学》第四版 P35）？

解：这是一个典型的配对设计，显然应当用配对设计差值的 t 检验来做。按照配对 t 检验对数据格式的要求，这里在输入数据时应当按照和上表相同的格式输入，即每个变量（一列）代表一个组，而每条记录（一行）代表一对数据。最终数据集中有 newdrug 和 placebo 两个变量，见数据文件 pair.sav。分析时注意变量需要成对选入。

配对号	新药组	安慰剂组	差值 d
1	4.4	6.2	-1.8
2	5.0	5.2	-0.2
3	5.8	5.5	0.3
4	4.6	5.0	-0.4
5	4.9	4.4	0.5
6	4.8	5.4	-0.6
7	6.0	5.0	1.0
8	5.9	6.4	-0.5
9	4.3	5.8	-1.5
10	5.1	6.2	-1.1

Analyze→Compare Means→Paired-Samples T Test

Paired Variables 框: newdrug、placebo

先后单击变量名即可成对选入

OK

12.4.2 界面说明

【主对话框】(见图12.7)

1. Paired Variable 框: 用于选入希望进行比较的一对或几对变量——注意这里的量词是对而不是个。选入变量需要成对成对的选入, 依次选中两个成对变量, 再单击  将其选入。如果只选中一个变量, 则  按钮为灰色, 不可用。

2. Current Selections 信息栏: 用于动态反映当前所选中的变量名称。

【Options】子对话框

和前面完全相同, 此处不再重复。

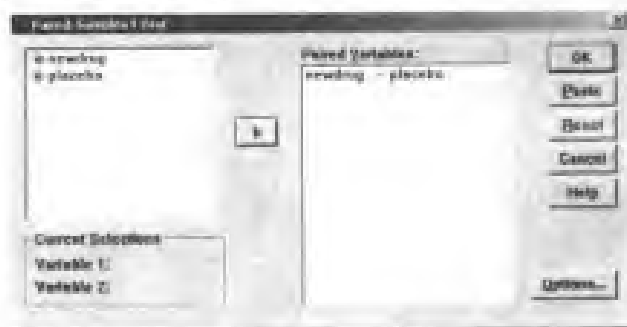


图 12.7 主对话框

12.4.3 结果解释

T-Test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	NEWDRUG	5.0800	10	.61988	.19596
	PLACEBO	5.5100	10	.64023	.20246

配对变量各自的统计描述, 此处只有1对, 故只有 Pair 1。

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	NEWDRUG & PLACEBO	10	.020	.956

此处进行配对变量间的相关性分析。等价于 Analyze→Correlate→Bivariate, 详见回归分析一章。

Paired Samples Test

Paired Differences									
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
					Lower	Upper			
Pair 1	INEWDRUG - PLACEBO	-.4300	.8820	.2789	-1.0609	.2009	-1.542	9	.158

配对 t 检验表, 给出最终的检验结果, 由上表可见 $P=0.158$, 故尚不能认为该新药会影响女性血清总胆固醇含量。

上表的标题内容翻译如下:

		对子间的差异					t 值	自由度	P 值 (双侧)
		差值均数	标准差	标准误	均数的 95%可信区间				
					下限	上限			
第一对	NEWDRUG - PLACEBO	-.4300	.88198	.27891	-1.0609	.2009	-1.542	9	.158

有兴趣的朋友可以将该数据变换格式, 采用单样本差值 t 检验或者配伍设计方差分析的方法做一下, 看看输出有无不同。

12.5 One-Way ANOVA 过程

One-Way ANOVA 过程用于进行两组及多组间样本均数的比较, 即成组设计的方差分析。如果做了相应选择, 还可进行随后的两两比较, 甚至于精确设定均数比较方式。在本章的内容中它是最为复杂的一个, 但是有了前面的基础, 拿下它应该不成问题。



为什么多个样本均数的比较不能用两两 t 检验来进行呢?

统计学的结论都是概率性的, 假设实际情况是 H_0 成立, 那么根据我们设定的 α (比如 $\alpha=0.05$) 水准, 平均每 100 次检验中有五次会得出拒绝 H_0 的错误结论。如果 k 个样本均数进行比较时采用两两 t 检验, 则共需作 $k!/2!(k-2)!$ 次比较, 每次比较不犯第一类错误的概率为 $(1-0.05)=0.95$, 当这些检验独立进行时, 则每次比较均不犯第一类错误的概率为 $0.95^{k!/2!(k-2)!}$, 相应的犯第一类错误的概率为 $1-0.95^{k!/2!(k-2)!}$, 远远大于设定的 0.05。例如 5 个样本均数作比较, 总的 α 水准就变成了 0.4013。就好像考试及格线原本是 60 分, 现在被降到了 20 分, 导致考试的权威性大打折扣一样。因此, 多个均数比较时不宜采用 t 检验作两两比较。

方差分析是基于变异分解的原理进行的，在单因素方差分析中，整个样本的变异可以看成由如下两个部份构成：

$$\text{总变异} = \text{随机变异} + \text{处理因素导致的变异}$$

其中随机变异是永远存在的，处理因素导致的变异是否存在就是我们希望研究的目标。即只要能证明它不等于 0，就等同于证明了处理因素的确存在影响。

那么，这一等式中的各项能否量化？首先考虑各组内部的变异（组内变异），该变异只反映随机变异的大小；再看各组均数的差异（组间变异），它反映了随机变异的影响与可能存在的处理因素的影响之和。如果用方差表示变异大小，则有如下等式成立：

$$\text{总变异} = \text{组内变异} + \text{组间变异}$$

并且该等式和上面的等式存在着如下的对应关系：

$$\begin{array}{ccc} \text{总变异} = \text{随机变异} + \text{处理因素导致的变异} & & \\ \downarrow \quad \downarrow \quad \searrow & & \\ \text{总变异} = \text{组内变异} + \text{组间变异} & & \end{array}$$

这样，我们可采用一定的方法来比较组内变异和组间变异的大小，如果后者远远大于前者，则说明处理因素的影响的确存在，如果两者相差无几，则说明该影响不存在，以上就是方差分析法的基本思想。

方差分析本身是完美的，但在解决实际问题的时候，我们往往仍需要回答多个均数间究竟是哪些和哪些存在差异，这样问题又回到了两两比较上。针对两两比较时如何控制一类错误的大小，统计学上已经发展出了一系列的两两比较方法，用于方差分析后进一步的检验。

和 t 检验时的情况类似，方差分析也要求各样本来自正态总体，且各总体方差相等。如这些条件不满足，则应进行变量变换，或放弃使用该方法。

12.5.1 引例

例 12.5 某职业病防治院对 31 名石棉矿工中的石棉肺患者、可疑患者及非患者进行了用力肺活量 (L) 测定，问三组石棉矿工的用力肺活量有无差别（杨树勤，《卫生统计学》第三版 P44）？

石棉肺患者	1.80	1.40	1.50	2.10	1.90	1.70	1.80	1.90	1.80	1.80	2.00
可疑患者	2.30	2.10	2.10	2.10	2.60	2.50	2.30	2.40	2.40		
非患者	2.90	3.20	2.70	2.80	2.70	3.00	3.40	3.00	3.40	3.30	3.50

解：数据为定量资料，设计为成组设计，由于是三组均数的比较，应当采用单因素方差分析。首先需要按照方差分析的格式要求输入数据，见数据集 lung.sav。其中分组变量为 group，用于记录该矿工属于哪一组，三组取值分别为 1、2、3；结果变量为 lung，用于记录矿工的肺活量值。由于方差分析要求方差齐等条件，分析时应当同时作方差齐性检验。同时为了回答实际问题，我们随后又进行了两两比较，方法为最常用的 SNK 法。

Analyze → Compare Means → One-Way ANOVA

Dependent List 框: lung

要分析的结果变量为 lung

Factor 框: group

分组变量为 group

Options:

☒ Homogeneity-of-variance

要求进行方差齐性检验

Continue

Post Hoc: ☒ S-N-K: Continue

两两比较方法采用 SNK 法

OK

12.5.2 界面说明

【主对话框】(见图 12.8)

1. Dependent List 框: 选入需要分析的变量, 如果选入多个结果变量(应变量), 则系统会依次对其进行单因素方差分析。

2. Factor 框: 选入需要比较的分组因素, 只能选入一个。

【Contrast 子对话框】(见图 12.9)

该对话框有两个用途: 对均数的变动趋势进行趋势检验; 定义根据研究目的需要进行的某些精确两两比较。由于该对话框太专业, 也较少用, 这里只做简单介绍, 在综合实例中会结合具体例题讲解。

1. ☐ Polynomial: 定义是否在方差分析中进行趋势检验, 即随着组别的变化, 各组均数是否呈现某种变化趋势。

2. Degree 下拉列表: 和 Polynomial 复选框配合使用, 用于定义需检验的趋势曲线的最高次方项, 可选择从线性趋势一直到五次方曲线。如果你选择了高次方曲线, 系统会给出所有相应各低次方曲线的拟合优度检验结果(比如选择 3 次方曲线时, 系统会给出线性、二次方、三次方三个结果), 以供你选择。

3. Coefficients 框: 精确定义某些组间均数的比较。这里按照分组变量升序给每组一个系数值, 注意最终所有系数值相加应为 0。比如说在上例中要对第一、三组进行单独比较, 则在这里给三组分配系数为 1、0、-1, 就会在结果中给出相应的检验内容。



图 12.8 主对话框



图 12.9 Contrast 子对话框



所有系数值相加不为 0 时仍可检验, 但 SPSS 不推荐这样做! 因为此时该检验的通用条件已被违反, 其结果的准确性可疑, 分析结论仅供参考。在 SPSS 的帮助中对此有明确的说明。

4. Coefficients Total 信息栏：动态提供键入系数的总和，以免用户因疏忽而导致系数总和不为 0。

【Post Hoc】子对话框（见图 12.10）

用于选择进行各组间两两比较的方法。

1. Equal Variances Assumed 复选框组：提供了一些当各组方差齐时可用的两两比较方法，共有 14 种，这里不一一列出了，只解释最常用的一些如下，除最后的 Dunnett 法外，其余几种大致是按从最敏感到最保守的顺序排列：

◇ LSD：即 LSD 法，实际上就是 t 检验的变形，只是在变异和自由度的计算上利用了整个样本信息，而不仅仅是所比较两组的信息。因此它敏感度最高，在比较时仍然存在放大 α 水准（一类错误）的问题，但换言之就是总的二类错误非常的小，要是 LSD 法都没检验出差别，那恐怕是真的没差别。

可能有的朋友对 LSD 法竟然就是 t 检验的变形一事感到不可思议，这种仍然存在缺陷的方法怎么还能允许使用？！别急，LSD 法有着特殊的应用目的，它只针对已计划好的某两个或几个组间的比较来使用（详后），也就是说，根据我们的分析目的，多重比较时校正 α 水准的原则并非是要死抱着不放。对这个问题我不准备深入探讨，只在此引用 Rothman 描述的死守多重比较校正原则时推出的荒谬性结论：如果赞同始终都校正，那么是否应对一篇论文中所评价的比较次数，或者是一系列分析相同数据的论文所评价的比较次数，甚至是整个研究生涯所进行的比较次数都来个校正吗？

◇ S-N-K：即 Student Newman Keuls 法，是运用最广泛的一种两两比较方法。它采用 Student Range 分布进行所有各组均值间的配对比较。该方法保证在 H_0 真正成立时总的 α 水准等于实际设定值，即控制了一类错误。

◇ Bonferroni：由 LSD 法修正而来，通过设置每个检验的 α 水准来控制总的 α 水准，该方法的敏感度介于 LSD 法和 Scheffe 法之间。

◇ Sidak：也是从 t 检验修正而来，和 Bonferroni 法非常相似，但比 Bonferroni 法保守。

◇ TUKEY：即 Tukey's honestly significant difference 法（Tukey's HSD），同样采用 Student-Range 统计量进行所有组间均值的两两比较。但与 S-N-K 法不同的是，它控制的是所有比较中最大的一类错误概率值不超过 α 水准。

◇ Scheffe：当各组人数不相等，或者想进行复杂的比较时，用此法较为稳妥。它检验的是各个均数的线性组合，而不是只检验某一对均数间的差异，并控制整体 α 水准等于 0.05。但正因如此，它相对比较保守，有时候方差分析 F 值有显著性，用该法两两比较却找不出差异来。

◇ Dunnett：将所有的处理组均数分别与指定的对照组均数进行比较，并控制所有比较中最大的一类错误概率值不超过 α 水准，请注意该方法并不适用于完全两两比较的情况。选定此方法后会激活下面的 Control Category 框，用于设定对照组及单双侧检验。

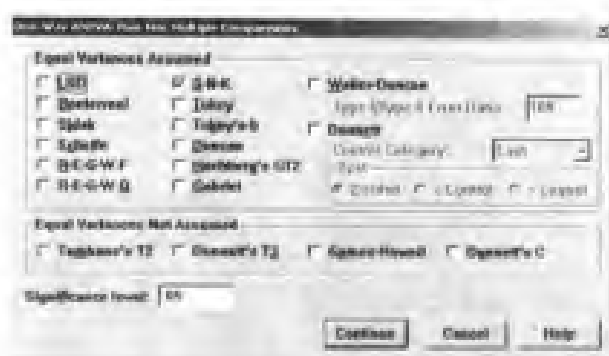


图 12.10 Post Hoc 子对话框

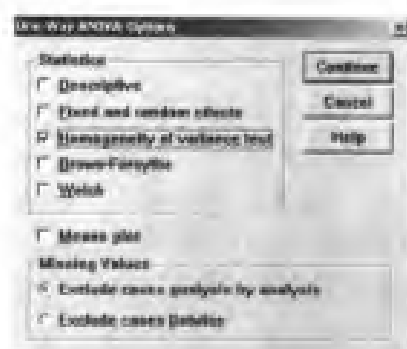


图 12.11 Options 子对话框

2. Equal Variances Not Assumed 复选框组：提供了方差不齐时可以采用的两两比较方法，共有四种可以选择，一般认为是 Games-Howell 法稍好一些，推荐使用。不过在我看来，由于这方面统计学界尚无定论，建议大家最好在方差不齐时直接使用非参数检验方法，具体的非参数两两比较方法会在相应章节中讲述。

如何在如此之多的两两比较方法中选出合适的一种是个令人头痛的问题。以前国内外都以 SNK 法最为常用，但根据研究，当两两比较的次数极多时，该方法的假阳性非常之高，最终可以达到 100%！因此比较次数较多时，包括 SPSS 和 SAS 在内的权威统计软件都不再推荐使用此法。

根据对相关研究的检索结果，除了参照所研究领域的惯例外，一般可以参照如下标准：如果存在明确的对照组，要进行的是验证性研究，即计划好的某两个或几个组间（和对照组）的比较，宜用 Bonferroni（LSD）法；若需要进行的是多个均值间的两两比较（探索性研究），且各组人数相等，适宜用 Tukey 法；其他情况宜用 Scheffe 法。该标准仅供大家参考。

3. Significance Level 框：定义两两比较时的显著性水平，默认是 0.05，一般来说不用更改。

【Options 子对话框】（见图 12.11）

1. Statistics 复选框组：提供了一些可选的统计指标，请注意它们并非可有可无。

- ✧ Descriptive：为各组输出常用统计描述指标，如均值、标准差等。
- ✧ Fixed and random effects：为 11.0 版新增，按固定效应模型输出标准差、标准误和 95%可信区间，同时按随机效应模型输出标准误、95%可信区间和成分间方差。关于这两种模型的详情请参见一般线性模型部分。
- ✧ Homogeneity-of-variance：方差齐性检验，这是常常被人忽略的一项重要功能。
- ✧ Brown-Forsythe：为 11.0 版新增，采用 Brown-Forsythe 统计量检验各组均值是否相等，当方差不齐时，该方法要比方差分析更为稳健。
- ✧ Welch：为 11.0 版新增，采用 Welch 统计量检验各组均值是否相等，当方差不齐时，该方法要比方差分析更为稳健。

2. Means plot：用各组均值做图，同时可辅助对均值间趋势做出判断。

3. Missing Values 单选框组：定义分析中对缺失值的处理方法，内容与前面几个过程相同，不再赘述。

12.5.3 结果解释

Oneway

Test of Homogeneity of Variances

肺活量			
Levene Statistic	df1	df2	Sig.
2.852	2	28	.075

方差齐性检验结果，Levene 统计量为 2.852，在当前自由度下对应的 P 值为 0.075，因此可以认为样本所在各总体的方差齐。

ANOVA

肺活量					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	9.266	2	4.633	84.544	.000
Within Groups	1.534	28	5.480E-02		
Total	10.800	30			

这实际上是一个典型的方差分析表。给出了单因素方差分析的结果，可见 $F=84.544$ ， $P<0.001$ 。因此可认为三组矿工用力肺活量不全相同。上表的标题内容翻译如下：

	离均差平方和 SS	自由度	均方 MS	F 值	P 值
组间变异	9.266	2	4.633	84.544	.000
组内变异	1.534	28	5.480E-02		
总变异	10.800	30			

下面将要输出的是两两比较的结果。

Post Hoc Tests

Homogeneous Subsets

肺活量

Student-Newman-Keuls ^{a,b}				
Subset for alpha = .05				
分组情况	N	1	2	3
石棉肺患者	11	1.7909		
可疑患者	9		2.3111	
非患者	11			3.0818
Sig.		1.000	1.000	1.000

Means for groups in homogeneous subsets are displayed.

- a. Uses Harmonic Mean Sample Size = 10.241.
b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

上表是用 S-N-K 法进行两两比较的结果，简单的说，在表格的纵向上各组均数按

大小排序，然后在表格的横向上被分成了若干个亚组，不同亚组间的 P 值小于 0.05，而同一亚组内的各组均数比较的 P 值则大于 0.05。从表中可见，石棉肺患者、可疑患者和非患者被分在了三个不同的亚组中，因此三组间两两比较均有差异；由于各个亚组均只有 1 个组别进入，因此最下方的组内两两比较 P 值均为 1.000（自己和自己比较，当然绝对不会有差异了）。



从上面的解释大家可得：SPSS 在用 SNK 法进行两两比较时，如果有差异，则只会告诉你 P 值小于预定的界值（默认为 0.05），而不会给出具体的 P 值有多大。

12.5.4 进一步分析的结果

【均数图】

图形往往可以提供更加直观的信息，为此我们用 Options 子对话框中的 Means plot 复选框做出三组的均数图如图 12.12 所示。

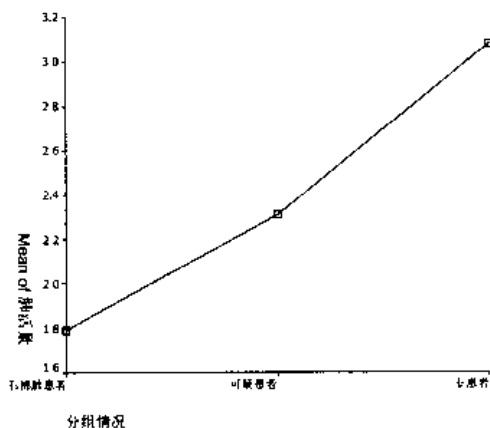


图 12.12 均数图

这是根据样本均数做出的均数图，可以直观了解结果变量随自变量的变化情况，可见数据明显有患病情况越轻，肺活量越大的趋势。这进一步肯定了两者的关系。

【用 LSD 法进行两两比较】

上面我们采用的是 SNK 法来进行了两两比较，如果希望以非患者作为对照，来研究石棉肺患者及可疑患者的肺活量有无降低，则应当采用 LSD 法，结果如下：

Multiple Comparisons

Dependent Variable: 肺活量

LSD

(I) 分组情况	(J) 分组情况	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
石棉肺患者	可疑患者	-.5202*	.10522	.000	-.7357	-.3047
	非患者	-1.2909*	.09982	.000	-1.4954	-1.0864
可疑患者	石棉肺患者	.5202*	.10522	.000	.3047	.7357
	非患者	.7707*	.10522	.000	.5552	.9862
非患者	石棉肺患者	1.2909*	.09982	.000	1.0864	1.4954
	可疑患者	.7707*	.10522	.000	.5552	.9862

* The mean difference is significant at the .05 level.

由于 LSD 法需要有一个对照组, 分析结果中就将所有组依次作为对照组, 让其余各组和它进行比较, 根据分析目的, 我们应当看最下面以非患者组作为对照的结果。表格中依次给出的是两组间均数差值、差值的标准误、P 值以及差值的可信区间。其中如果均数差值有统计学意义, 则自动在后面加上“*”作为标记。由上表可见石棉肺患者及可疑患者的肺活量与正常人的肺活量差值有统计学意义, 由于差值均为正, 因此他们都低于正常人的肺活量。

12.6 综合分析实例

12.6.1 两样本 t 检验: 巧用 Cutpoint

例 12.6 今欲研究中风病人入院时状况与发病——入院时间的关系, 共收集了 88 例中风病人的发病——入院时间与入院时的 GCS 评分, 见数据文件 gcs.sav, 变量 time 为就诊时间分级, 1 级为 5 小时内, 级别越高, 时间越久; gcs 为入院时的 GCS 评分。请问及时入院的病人入院时 gcs 评分与其他人有无不同。

解: 做一频数表即可知 time 共分五级, 但 4、5 两级人数非常少, 这种情况不宜直接进行方差分析, 可以将这两级去除, 或与 3 级合并进行分析。题中关心的是及时入院者与其他人有无不同, 因此最简单的做法是将 2~5 级合并, 与第 1 级进行 t 检验。我们可以用 Compute 过程生成一个新的分组变量, 但两样本 t 检验过程中的 Cut point 框可以使直接完成该分析。

Analyze→Compare Means→Independent-Samples T Test

Test Variable(s) 框: gcs

Grouping Variable 框: time

选中变量 group: **Define Groups:**

☒ Cut point 框: 键入 2

Continue

OK

要分析的变量为 gcs

分组变量为 time

定义检验的两组

Time<2 的和 ≥2 的进行比较

分析结果中的分组统计表如下:

Group Statistics				
	就诊时间分级	N	Mean	Std. Deviation
GCS 评分	≥ 2.00	54	11.7963	2.2851
	< 2.00	34	11.1176	2.3454
				Std. Error Mean
				.3110
				.4022

从上表中可以看到所有记录按 time 的取值是否小于 2 被分为了两组进行比较。下面的 t 检验分析结果略。

12.6.2 方差分析: 指定均数的比较

例 12.7 今欲研究某种毒素 (T2) 对小白鼠的毒性作用, 共分 0、0.125、0.25、0.5、1 (mg/L) 五个浓度组, 每组 6 只小白鼠, 结果如下表。根据前期实验, 已知 0.5mg/L 以下浓度不会对结果造成影响, 请根据这个实验的结果判断 1mg/L 的浓度是否对结果有影响?

0 mg/L 组	0.125 mg/L 组	0.25 mg/L 组	0.5 mg/L 组	1 mg/L 组
.362	.356	.366	.351	.262
.365	.357	.362	.359	.271
.362	.358	.364	.358	.271
.295	.288	.294	.285	.200
.290	.293	.287	.288	.198
.291	.289	.292	.285	.205

解：该题为一典型的单因素方差分析问题，但现在特殊的是已有关于实验的相关信息存在，其前四组应当是没有差异的，我们重点关心的是第五组和前四组有无差异，这用普通的两两比较会得出四个结论，可能会造成混乱。而用 **Contrast** 子对话框恰恰就可以轻松完成这种需要精确定义的比较。数据已输入为数据文件 T2.sav，其中变量 T2 为结果变量，nongdu 为分组变量。

Analyze→Compare Means→One-Way ANOVA

Dependent List 框: t2

要分析的结果变量为 t2

Factor 框: nongdu

分组变量为 nongdu

Options:

☒ Homogeneity-of-variance

要求进行方差齐性检验

☒ Means plot

要求做出均数图

Continue

Contrasts:

Coefficients 框: 1: Add: 1: Add

定义 nongdu=1、2 两组的系数为 1

Coefficients 框: 1: Add: 1: Add

定义 nongdu=3、4 两组的系数为 1

Coefficients 框: -4: Add

1mg/L 组的系数为-4，确保总和为 0

Continue

OK

结果如下所示：

Test of Homogeneity of Variances

结果指标

Levene Statistic	df1	df2	Sig.
1.314	4	25	.292

方差齐性检验结果，显示各样本所在总体的方差齐。

ANOVA

结果指标

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3.939E-02	4	9.847E-03	6.732	.001
Within Groups	3.657E-02	25	1.463E-03		
Total	7.595E-02	29			

方差分析的结果显示浓度的确对结果有影响。

Contrast Coefficients

Contrast	浓度(mg/L)				
	0	0.125	0.25	0.5	1
1	1	1	1	1	-4

上表为 Contrast 语句中各组设定的系数列表，主要用于备查。

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
结果指标	Assume equal variances	1	.36150	6.9825E-02	5.177	25	.000
	Does not assume equal	1	.36150	6.6012E-02	5.315	7.953	.001

上表为按 Contrast 语句设定进行的 1mg/L 组与前四组均数比较的结果，会输出方差齐和方差不齐两种情况下的 t 值、自由度和 P 值，根据前面方差齐性检验的结果，应当看方差齐的一行，从中可见差异的确有统计学意义。



对统计比较熟悉的朋友可能会有疑问：这里不是也可以将前四组合并，当作一组来和第五组进行 t 检验吗？那么这样的 t 检验和现在的做法有什么区别？实际上，现在的做法和 t 检验没有本质的区别，不信可以对照一下方差不齐时的 t 值和 P 值，应当非常接近。但是，现在的做法充分利用了样本信息，同时对变异做了更为精细的分解，要比直接作 t 检验更好。

Means Plots

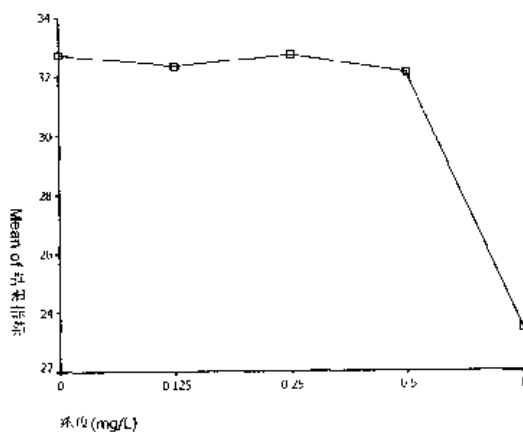


图 12.13 均数图

图 12.13 为 5 组的均数分布图，可见前四组的均数非常接近，第五组才迅速下降，和以前实验及本次检验得出的结论相一致。

12.6.3 方差分析：均数间曲线趋势的判断

例 12.8 今欲研究胸腺增生病人中增生情况与 titinab 值的关系，共调查了 141 名病人，数据见 titinab.sav。变量 class 表示胸腺增生情况，共分为 1~5 级；变量 titinab 为 titinab 测量值。希望能找出 titinab 随增生级别提高的变化趋势。

解：本例实际上是五组均数的比较，用单因素方差分析即可，但现在实际上已经

明确两者有关系，重点在于找出 titnab 随增生级别的变化趋势（几次方曲线？），用 **Contrast** 子对话框中的趋势检验即可完成，最后可结合均数图做出判断。本例共有五组，最多可拟合四次方曲线，但这显然不实用，我们只拟合 3 次方曲线即可。

在进行方差分析之前，进行了变量的描述，发现 titnab 呈正偏态分布，因此对其做了对数变换，变量描述的输出结果从略。

Transform→Compute

Target Variable 框: logtitin

新变量名为 logtitin

Numeric Expression 框: lg10(titinab)

新变量等于 titinab 的对数

OK

Analyze→Compare Means→One-Way ANOVA

Dependent List 框: logtitin

要分析的结果变量为 logtitin

Factor 框: class

分组变量为 class

Options:

☒ Homogeneity-of-variance

要求进行方差齐性检验

☒ Means plot

要求做出均数图

Continue

Contrasts:

☒ Polynomial

要求进行拟合优度检验

Degree 框: Cubic

需检验的最高次曲线为三次方曲线

Continue

OK

主要的分析结果如下：

Oneway

Test of Homogeneity of Variances

LOGTITIN

Levene Statistic	df1	df2	Sig.
.577	4	136	.680

方差齐性检验的结果显示各组间的方差齐。

ANOVA

LOGTITIN

			Sum of Squares	df	Mean Square	F	Sig.
Between Groups	(Combined)		1.171	4	.293	11.189	.000
	Linear	Unweighted	.714	1	.714	27.295	.000
	Term	Weighted	1.017	1	1.017	38.857	.000
		Deviation	.154	3	.051	1.967	.122
	Quadratic	Unweighted	.136	1	.136	5.202	.024
	Term	Weighted	.154	1	.154	5.893	.017
		Deviation	.000	2	.000	.004	.996
	Cubic	Unweighted	.000	1	.000	.006	.941
	Term	Weighted	.000	1	.000	.005	.942
		Deviation	.000	1	.000	.003	.959
Within Groups			3.559	136	.026		
Total			4.731	140			

上表实际上仍为方差分析表，但加入了曲线拟合情况的检验。对于每种曲线，SPSS 都会给出考虑和不考虑各组记录数权重的多重比较结果，以及考虑权重时离差的检验结果。多重比较结果以最前面的一项为准即可，从上表可见，各组均数明显不呈线性关系，也拒绝了二次方曲线的假设，只有三次方曲线的拟合度非常好，残差也无统计意义。

如果按照一般的分析思路，我们到此为止已经得出了最终的分析结论——随着增生级别的增高，logitnab 呈三次曲线的上升趋势，似乎问题已被解决。但是，考虑到我们一共只有五个均数，得出的结论是三次方曲线的话是不是阶次太高了？要知道在统计分析中的一个重要原则是简洁就是美，过于复杂的模型不仅难于使用，而且有可能导致错误的分析结果。为了慎重起见，下面我们使用均数图来直观的了解一下均数间的变动趋势。

Means Plots

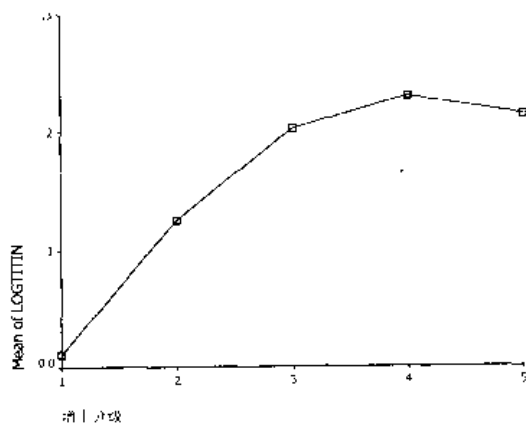


图 12.14 均数图

图 12.14 为判断均数间趋势时非常重要的均数图，从图中可见各组均数呈较为明显的二次方曲线趋势，虽然前面的检验 P 值为 0.02，拒绝了该假设，但考虑到样本量较大，我们最终仍可以认为随着增生级别的增高，logitnab 呈二次曲线的上升趋势，并在 4 级时达到最高峰。请注意，**曲线形式的判定并不是完全靠 P 值来决定的**，许多时候要依靠图形以及专业知识来做出结论。