多元分析

典型相关分析CCA
  定义
  思路和关键步骤
    思路
    关键步骤
      数据分布假设
      相关性检验（构造似然比统计量）
      确定典型变量的组数
      标准化
      典型载荷分析
      计算前n个变量对反差的总贡献
      冗余分析
      案例
      典型变量系数
  $Python$
相关回归分析
方差分析$ANONA$
  检验
    分类
    正态性检验
      频数分布直方图
      正态$Q-Q$图
      分位数
      原理
      正态$P-P$图
      去趋势正态$P-P$图
      $K-S$检验
      总结

方差齐性检验
$Levene$检验
方差分析
单因素方差分析
步骤
事后多重比较
结果分析
描述
方差齐性检验
F检验结果
韦尔奇检验
事后比较结果
双因素方差分析
分类
多因素方差分析
假设检验
贝叶斯

典型相关分析 CCA

定义

求解多个变量与多个变量之间的相关性时，将它们转化为两个具有代表性的变量的相关。这个代表能较为综合、全面的衡量所在组的内在规律。

一组变量最简单的综合形式就是该组变量的线性组合。

目的：

1. 数据简化：用少量的线性组合来解释两组变量之间的相关作用；
2. 数据解释：寻找特征值，这些特征值对于解释两个变量集合之间的相互作用十分关键。

与主成分分析（PCA）：

1. 联系：都是线性分析，典型变量和主成分都是通过计算矩阵的特征值和特征向量得出的；

2. 区别：主成分分析中只涉及一组变量的相互依赖关系，而典型相关则扩展到了两组变量之间的相互依赖的关系之中，度量了这两组变量之间联系的强度。

思路和关键步骤

思路

假设两组原始变量为 $X_1, X_2....X_p$ ， $Y_1...Y_q$ ，由这两组变量生成的综合变量就是**典型变量**，表示为：

$$V_i = a_1X_1 + a_2X_2 + ...a_pX_p \iff a_iX_1$$
$$W_i = b_1Y_1 + ...b_qY_q \iff b_iX_2$$

典型相关性问题就是找到系数 $a_1...a_p$ 及 $b_1...b_q$ 使得 V ， W 相关性最大。

这种相关性是由**典型相关系数**决定的，由于特征值问题的特点，我们实际上找到的是多组典型变量 $(V_1, W_1)...(V_n, W_n)$ ，典型变量的**组数**由它们特征值贡献率来决定（和主成分一样）。

关键步骤

数据分布假设

两组数据服从联合正态分布；

相关性检验（构造似然比统计量）

如果两个随机变量互不相关，及两组变量协方差矩阵 $cor(X_1, X_2) = 0$ 。但是也有可能得到协方差矩阵不为零。

我们就应该对协方差矩阵是否为零进行假设，即检验假设：

$$H_0 : \sum_{12} = 0, H_1 : \sum_{12} \neq 0$$

$p \leq 0.05$ 表示在 95% 的置信条件下拒绝原假设，即认为两组变量**有关**。

确定典型变量的组数

观察典型变量系数对应，如果前 n 组典型变量的贡献率已经达到一个比较大的值，就选取 n 组典型变量。

例如：

典型变量对	典型相关系数	特征值	Wilks	模型自由度	误差自由度	F	P
第1对	0.994	0.989	0.002	12	114.059	87.392	0.000***
第2对	0.878	0.771	0.195	6	88	18.526	0.000***
第3对	0.384	0.147	0.853	2	45	3.882	0.028**

注：***、**、*分别代表1%、5%、10%的显著性水平

第一二对典型变量的贡献率分别为 $51.9\%, 40.4\%$ 。

标准化

由于典型相关分析涉及两组变量，不同的变量之间往往会有不同的量纲以及数量级别。必须先对数据进行标准化变换处理，然后再做典型相关分析。

典型载荷分析

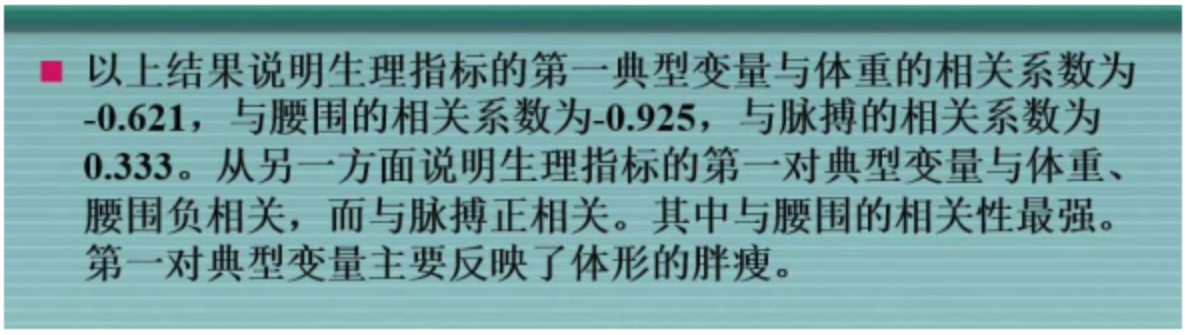
典型载荷是原始变量与典型变量之间的关系。

交叉载荷系数是典型变量与另一组变量各个变量的简单相关系数。

如：

集合 1 典型载荷

变量	1	2	3
体重x1	-.621	-.772	-.135
腰围x2	-.925	-.378	-.031
脉搏x3	.333	.041	.942



计算前 n 个变量对反差的总贡献

已解释的方差比例

典型变量	集合 1 * 自身	集合 1 * 集合 2	集合 2 * 自身	集合 2 * 集合 1
1	.451	.285	.408	.258
2	.247	.010	.434	.017
3	.302	.002	.157	.001

求得生理指标样本方差由自身 3 个典型变量解释的方差比例分别为：

第一典型变量解释的方差比例$=(0.621^2+0.925^2+0.333^2)/3=0.451$

第二典型变量解释的方差比例$=(0.772^2+0.377^2+0.041^2)/3=0.246$

第三典型变量解释的方差比例$=(0.135^2+0.031^2+0.942^2)/3=0.302$

冗余分析

进行样本典型相关分析时，分析**每组变量提取出的典型变量所能解释的该组样本总方差的比例**，称为冗余分析。

冗余分析包括**组内代表比例**和**交叉解释比例**，是典型相关分析中很重要的部分。

1. 组内代表比例：本组所有观测变量的总标准方差中由本组形成的各个典型变量所分别代表的比例；
2. 交叉解释比例：一组变量形成的典型变量对另一组观测变量的总标准方差所解释的比例， 是一种组间交叉共享比例。

作用：定量地测度典型变量所包含的原始信息量的大小。

案例

如果一个变量可以由另一个变量的方差来解释或者预测，那么就说这个方差部分与另一变量方差**冗余**。

典型变量系数

根据典型变量系数，可以得到典型变量的组成公式。

Python

参考材料：

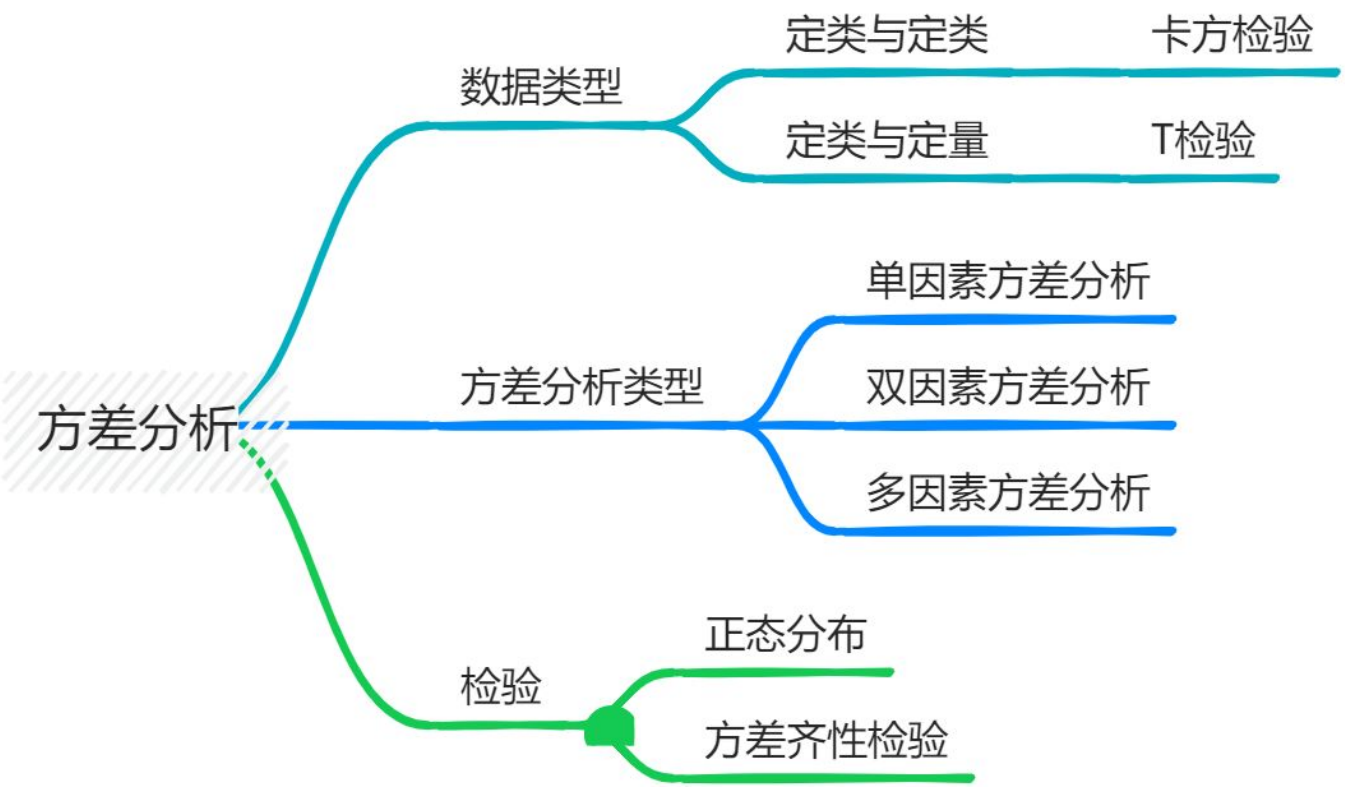
相关回归分析

方差分析 $ANONA$

方差分析又称 F 检验，用于定类数据与定量数据之间的关系情况，通过分析不同来源的变异对总变异的**贡献大小**，从而确定可控因素对研究结果影响力的大小。

适用范围：

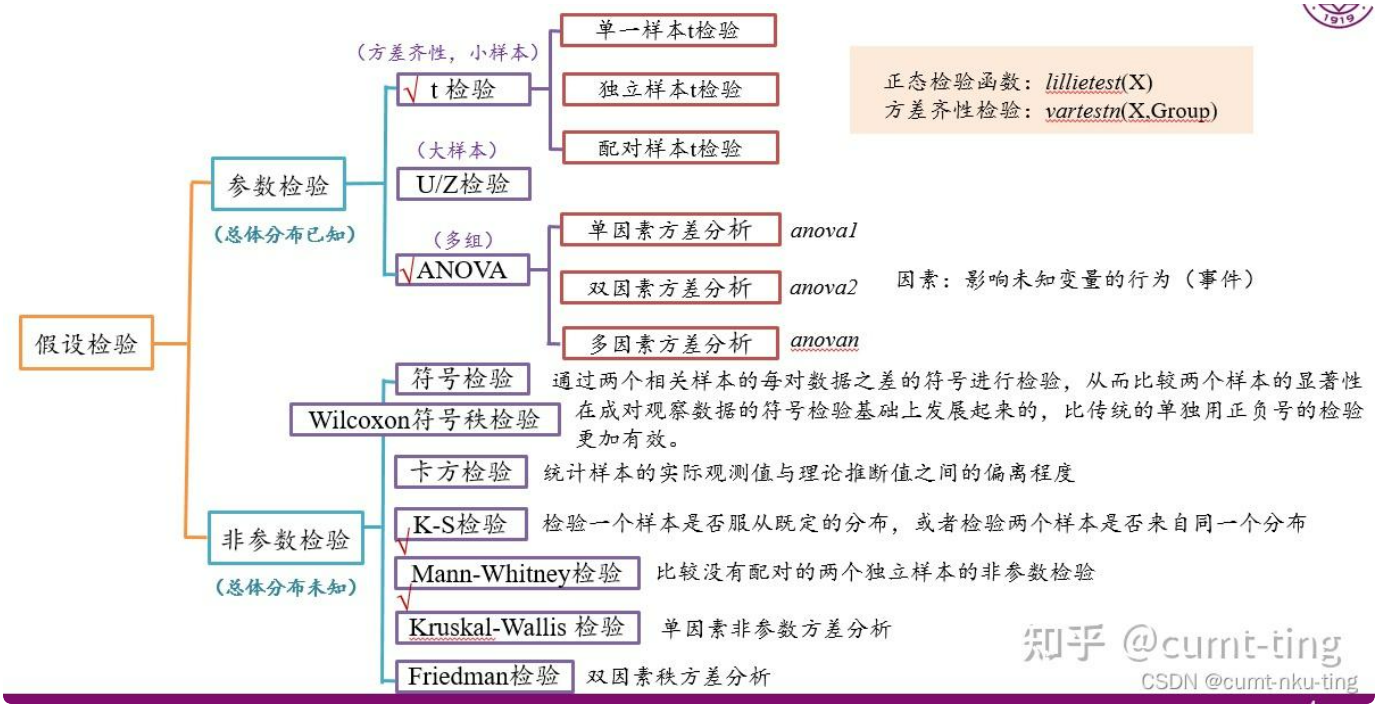
1. 自变量对因变量是否有显著影响；
2. 一个组内的多个样本，各组的平均值是否存在显著差异；



检验

分类

1. 参数检验：针对参数；
   非参数检验：针对整体分布情况；
2. 参数检验：利用到总体的信息（总体分布、总体的一些参数特征如方差），以**总体分布和样本信息**对总体参数作出推断；
   非参数检验：不需要利用总体的信息（总体分布、总体的一些参数特征如方差），以**样本信息**对总体分布作出推断；
3. 参数检验：只适用于等距数据和比例数据；
   非参数检验：主要用于记数数据，有时用作等距数据和比例数据，但**精确度**会降低。
4. 参数检验：对数据要求极其严格，必须保持数据连续性、分布性已知和总方差相等；
   非参数检验：检验效率较低。

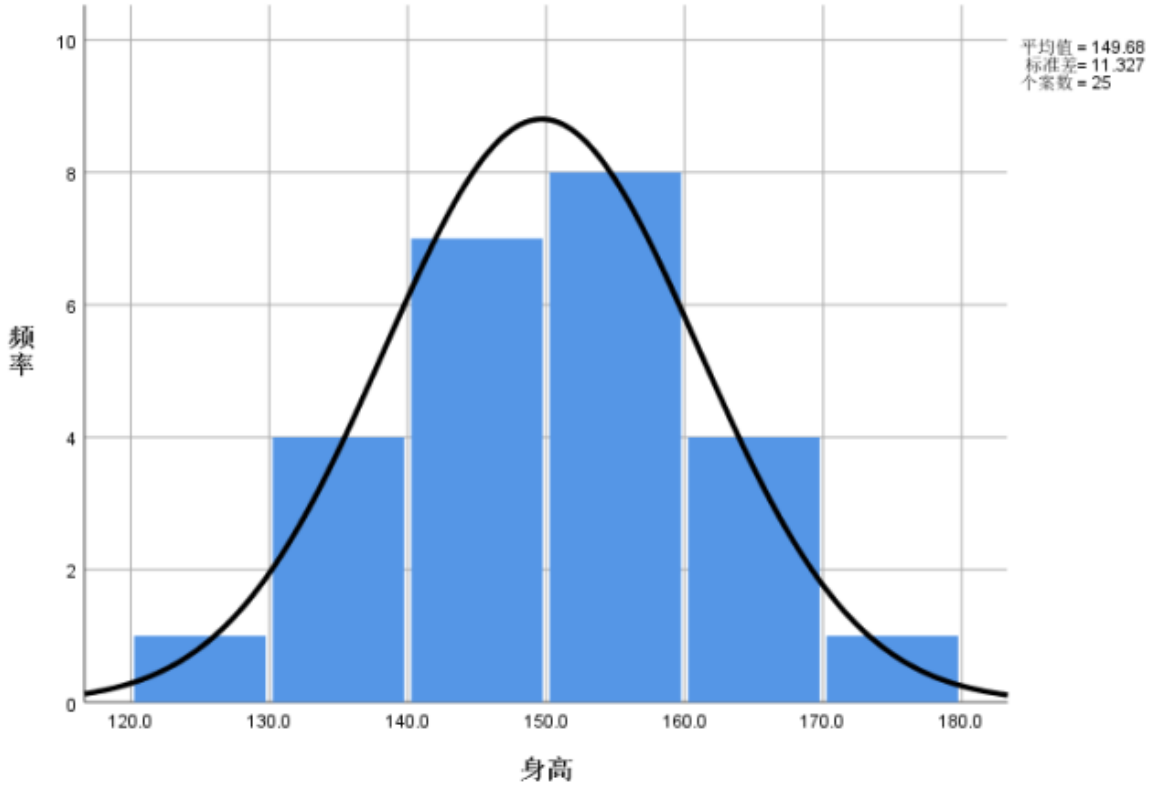


正态性检验

频数分布直方图

分析–描述性统计–频率

观察直方图的分布形状是否为一个倒扣“钟”型的对称形状，如果接近或相似，则可认为数据服从正态分布。

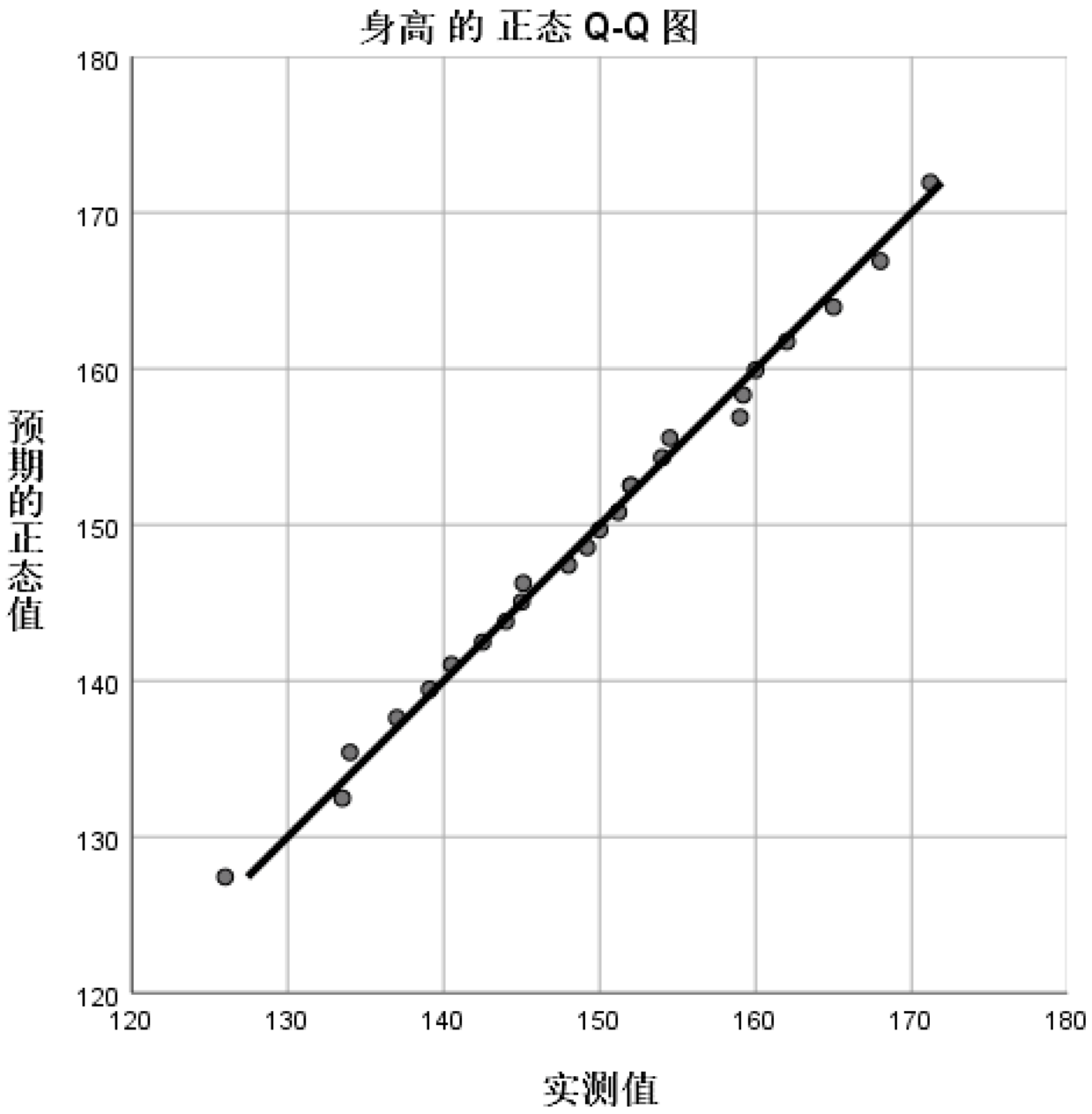


正态 $Q-Q$ 图

分析–描述性统计– $Q-Q$ 图。

$Q-Q$ 图（ Q 表示分位数）是一个概率图，以图形的方式比较两个概率分布，

如果 $Q-Q$ 图的点分布在一条直线上，分布在一条直线上则说明**近似或服从正态分布**。



分位数

设连续随机变量的函数为 $F(x)$ ，密度函数为 $f(x)$ ，对任意 $p \in (0,1)$ ，称满足条件

$$F(x_p) = \int_{-\infty}^{x_p} p(x)dx = p$$

x_p 为此分布的 p 分位数，又称为下侧 p 分位数。

原理

$Q-Q$ 图就是将一列样本数据的分位数与已知分布的一列数据的分位数相比较，从而来检验数据的分布情况。所以它的功能就是判断两列数据的分位数是否分布在 $y=x$ 的直线上。

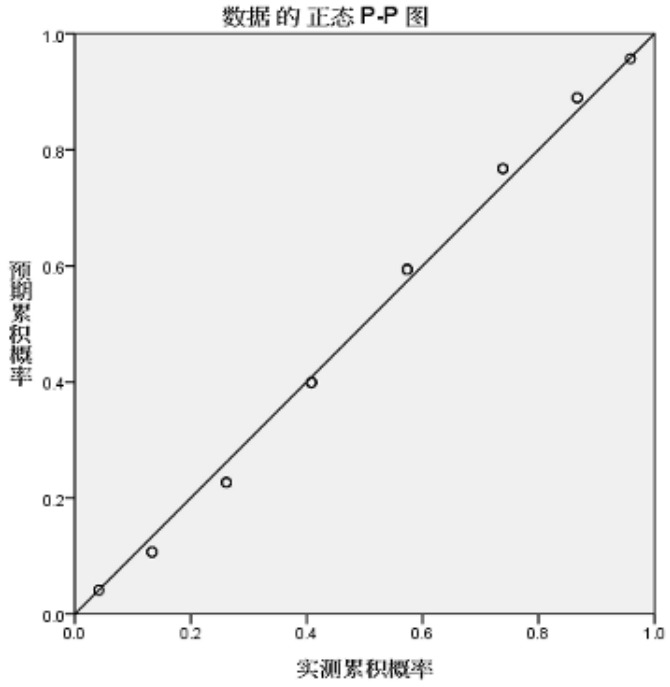
正态 $P-P$ 图

$P-P$ 图是显示数据概率的曲线图，中间的直线为**趋势线**，当数据愈加贴合趋势线的时候，说明数据的分布于理论的分布相差不大，进而认为数据基本符合**正态分布**。

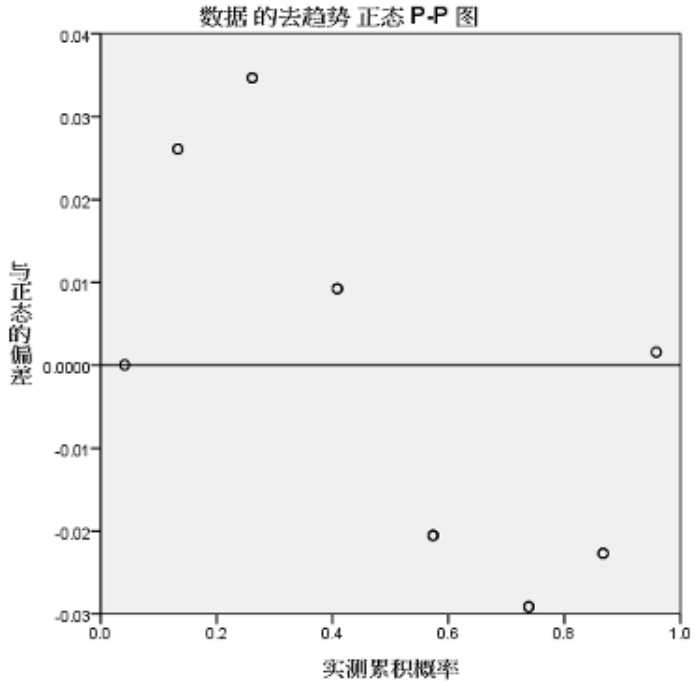
去趋势正态 $P-P$ 图

观察数据的趋势范围。

SPSS：分析，描述统计，P-P图。



p-p是显示数据概率的曲线图，中间的实线为趋势线，当数据愈加贴合趋势线的时候，说明数据的分布和和理论的分布差别不大，进而认为数据基本符合正态分布。



去趋势图，可以看出趋势的范围，上图大概在-0.03到0.04，这个图没什么用。

$K-S$ 检验

在 $K-S$ 检验中，如果显著性 $p \geq 0.05$ ，则认为是正态分布。

注：建议联合使用 $P-P$ 和 $K-S$ **图检验**表达数据的正态性，因为数据较多的时候， $K-S$ 图有可能小于0.05。

总结

综上，如果数据不满足正态分布，可以：

1. 进行对数处理。即对因变量进行转换，使其满足正态分布。
   需要注意的是转换后数据分析不容易解释；
2. 使用非参数检验。前面提到。参数检验的要求远远高于非参数检验，如果没有呈现出正态性特质，可使用非参数检验进行分析，但也要考虑性能的降低。

方差齐性检验

$Levene$ 检验

将每个值先转换为为该值与其组内均值的偏离程度，然后再用转换后的偏离程度去做方差分析，关于组内均值有多种计算方式，如平均数、中位数、截取平均数（去掉最大和最小值后求平均）。

方差齐性检验

		莱文统计	自由度 1	自由度 2	显著性
销售额	基于平均值	.040	1	196	.842
	基于中位数	.059	1	196	.809
	基于中位数并具有调整后自由度	.059	1	195.009	.809
	基于剪除后平均值	.036	1	196	.851

当四组样本的方差基本相同，且显著性 $p \geq 0.05$ 时，即数据符合方差齐性。

方差分析

单因素方差分析

用于检验单因素水平下的一个或多个独立因变量均值是否存在显著差异，即检验单因素各个水平的值是否来自同一个总体。

步骤

现在假定一个因素 A 具有 c 个水平的因变量进行方差分析。

1. 提出两种假设（原假设和备择假设）；

   H_0 ：各样本均值相同 $\mu_1 = \mu_2 ... = \mu_c$ ；

   H_1 ：各样本均值不全相同。

2. 计算各个样本的均值和方差；

$$\overline{x_j} = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$$

$\overline{x_j}$为第 j 个水平的样本均值

x_{ij}为第 j 个水平的第 i 各数值

n_j为第 j 各水平的样本容量

$$S_j^2 = \frac{\sum_{i=1}^{n_j}(x_{ij} - \overline{x_j})^2}{n_j - 1}$$

S_j^2是第 j 个水平的样本方差

3. 计算组间方差 MSB 和组内方差 MSE ；

组间方差记为 MSB，表示是 B 因素的均方

$$\text{MSB} = \frac{\sum_{j=1}^{c} n_j(\overline{x_j} - \bar{\bar{x}})^2}{c - 1}$$

$\sum_{j=1}^{c} n_j(\overline{x_j} - \bar{\bar{x}})^2$ 称为水平项平方和，记为 SSB；

$c - 1$是 SSB 的自由度；

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{c}\sum_{i=1}^{n_j} x_{ij}}{n_T}$$

$\bar{\bar{x}}$表示总的样本均值；

n_T表示每个样本容量之和；

组内方差记为 MSE，其计算公式为：

$$\text{MSE} = \frac{\sum_{j=1}^{c}\sum_{i=1}^{n_j}(x_{ij} - \overline{x_j})^2}{n_T - c}$$

$\sum_{j=1}^{c}\sum_{i=1}^{n_j}(x_{ij} - \overline{x_j})^2$ 称为误差项平方和，记为 SSE；

$n_T - c$是 SSE 的自由度；

4. 构造 F 型统计量进行检验；

   $F = MSB/MSE \sim F(c-1, nT-1)$

   如果假设 H_1 成立，则 $MSB > MSE$ ， F 大到某一临界值时，就可以拒绝 H_0 ，临界值的大小由给定的 α 和自由度决定。所以，当给定显著性水平为 α 时， F 的拒绝域为 $F > F\alpha(c-1, nT-c)$ 。

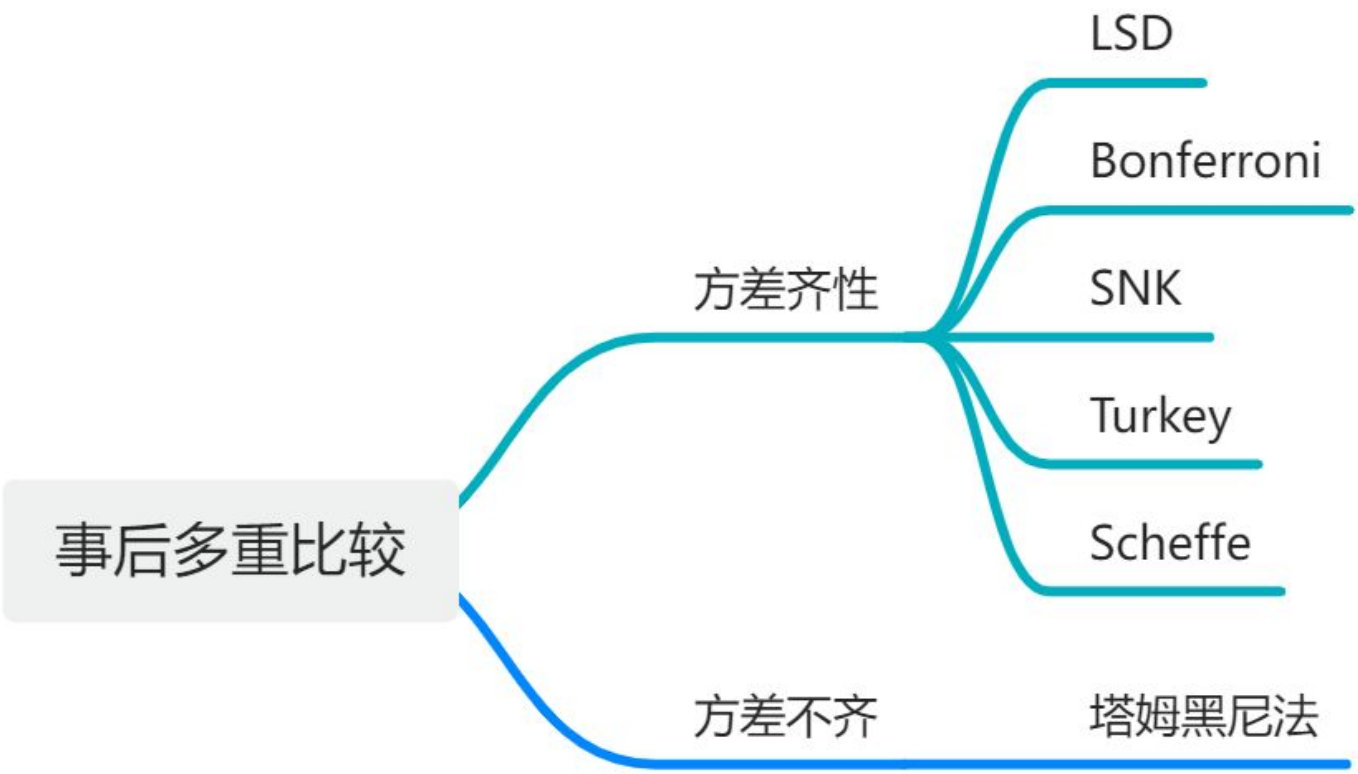
方差分析表				
方差来源	离差平方和	自由度*df*	均方*MS*	*F*值
组间	SSB	c-1	MSB	MSB/MSE
组内	SSE	n_T-1	MSE	
总方差	SST	n_T-c		

事后多重比较

指的是一个研究项目设计时未事先制定比较的组别和方法，而在统计分析阶段进行**任意**组别的均数两两比较。

在方差分析过程中，我们会选择一些检验方法。但在输出结果中，只有整体的F值，达到显著性水平，即 $p < 0.05$ 时，才会报告事后多重比较结果。

$SPSS$ 的事后多重比较分为**方差齐性**和**方差不齐**两种方式。



结果分析

描述

统计描述，包括均数、标准差、95%CI置信区间、最小值和最大值等。

描述

推铅球成绩

	个案数	平均值	标准 偏差	标准 错误	平均值的 95% 置信区间 下限	上限	最小值	最大值
A1教法	12	7.2033	1.09568	.31630	6.5072	7.8995	5.55	8.88
A2教法	10	6.2010	.94915	.30015	5.5220	6.8800	5.12	8.45
A3教法	11	5.9545	.62406	.18816	5.5353	6.3738	5.10	6.84
总计	33	6.4833	1.05032	.18284	6.1109	6.8558	5.10	8.88

方差齐性检验

方差齐性检验

		莱文统计	自由度 1	自由度 2	显著性
推铅球成绩	基于平均值	1.442	2	30	.252
	基于中位数	.924	2	30	.408
	基于中位数并具有调整后自由度	.924	2	24.290	.411
	基于剪除后平均值	1.417	2	30	.258

一般我们是选择第一行基于平均数的结果， $p \geq 0.05$ 时，方差齐性，可采取 F 检验。

F 检验结果

ANOVA

推铅球成绩

	平方和	自由度	均方	F	显著性
组间	10.094	2	5.047	6.006	.006
组内	25.208	30	.840		
总计	35.302	32			

韦尔奇检验

在方差不齐的时候才使用的检验。如果方差齐性，则不会出现检验结果。

平均值相等性稳健检验

推铅球成绩

	统计[a]	自由度 1	自由度 2	显著性
韦尔奇	5.585	2	18.670	.013

a. 渐近 F 分布。

事后比较结果

多重比较

因变量: 推铅球成绩

	(I) 教法	(J) 教法	平均值差值 (I-J)	标准 错误	显著性	95% 置信区间 下限	95% 置信区间 上限
邦弗伦尼	A1教法	A2教法	1.00233*	.39249	.048	.0071	1.9976
		A3教法	1.24879*	.38264	.008	.2785	2.2191
	A2教法	A1教法	-1.00233*	.39249	.048	-1.9976	-.0071
		A3教法	.24645	.40052	1.000	-.7692	1.2621
	A3教法	A1教法	-1.24879*	.38264	.008	-2.2191	-.2785
		A2教法	-.24645	.40052	1.000	-1.2621	.7692
塔姆黑尼	A1教法	A2教法	1.00233	.43604	.094	-.1337	2.1383
		A3教法	1.24879*	.36803	.010	.2789	2.2186
	A2教法	A1教法	-1.00233	.43604	.094	-2.1383	.1337
		A3教法	.24645	.35425	.873	-.7023	1.1952
	A3教法	A1教法	-1.24879*	.36803	.010	-2.2186	-.2789
		A2教法	-.24645	.35425	.873	-1.1952	.7023

*. 平均值差值的显著性水平为 0.05。

用来观察两两变量之间是否有显著性差异。

双因素方差分析

分析两个因素的不同水平对最终结果是否有显著影响。需要注意的是，双因素方差分析的结果仅适用于所收集数据的样本，不能推广到总体。

分类

1. 无交互作用：两个因素的效应之间相互独立，不存在相互关系。
2. 有交互作用：两个因素的结合会产生出一种新的效应。

多因素方差分析

假设检验

贝叶斯