



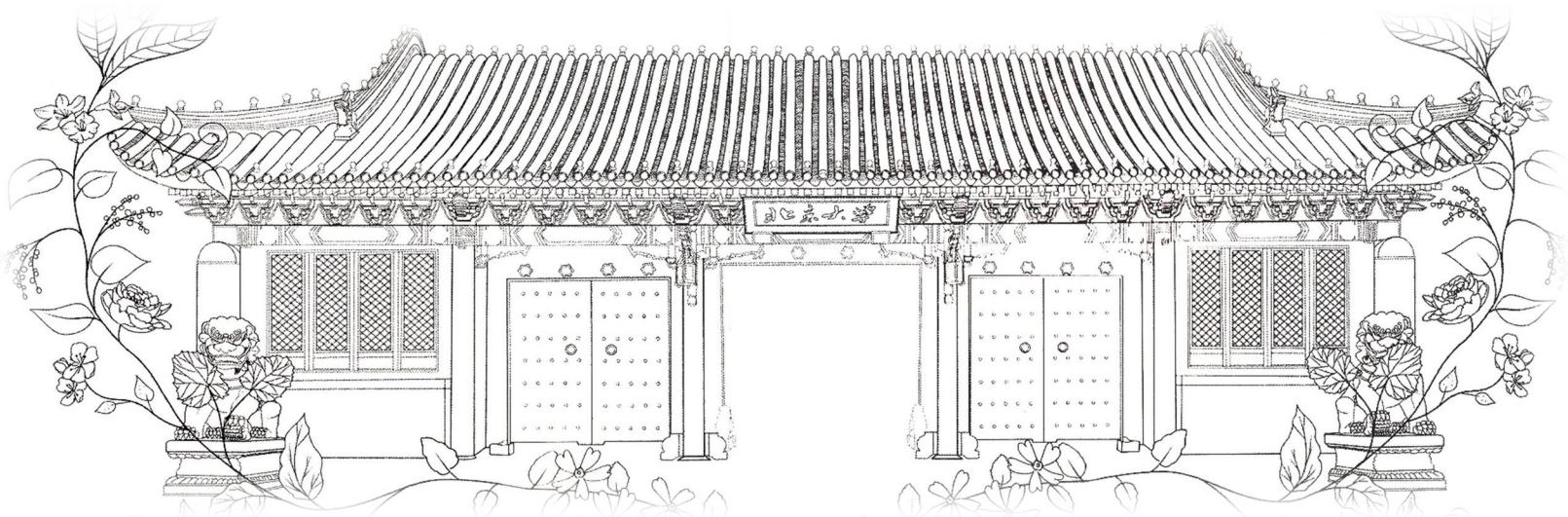
应用篇（第8-13周）

- 第8周：探索式数据分析
- 第9周：机器学习基础（作业3涉及）
- 第10周：时间序列分析（作业4涉及）
- 第11周：社会网络分析（作业4涉及）
- 第12周：文本数据分析（作业5涉及）
- 第13周：图像数据分析（作业5涉及）



《Python数据分析》

应用篇：探索式数据分析





探索式数据分析 (Exploratory Data Analysis, EDA)

- 含义

- 使用定量方法和可视化工具对数据集中变量的规律、趋势、隐含结构、离散情况、异常情况等做分析。

- 目标

- 理解数据的特征
- 识别潜在结构
- 抽取（识别）重要变量
- 发现异常值
- 决定最佳参数设定
-





探索式数据分析 vs. 验证式数据分析 (Confirmatory Data Analysis, CDA)

EDA	CDA
<ul style="list-style-type: none">• No hypothesis at first• Generate hypothesis• Uses graphical methods (mostly)	<ul style="list-style-type: none">• Start with hypothesis• Test the null hypothesis• Uses statistical models





目录

- 探索式数据分析 (EDA)：
 - 一个变量的分析
 - 两个变量的分析
 - 三个或三个以上变量的分析





一个变量的分析





一个变量的分析

- 描述性统计：最值、均值、百分位数、众数
- 分布
 - 表格或柱状图
 - 直方图
 - 密度分布
 - 累积分布
 - 顺便一提：正态分布如何检验？
- 对于同一个变量的多组样本
 - 参数检验
 - 非参数检验





描述性统计

- 最大值、最小值
- 众数
- 平均值
- 中位数和其他百分位数

```
# 平均数  
np.mean(data_set)  
np.mean(data_set["Lag1"])
```

```
# 中位数  
np.median(data_set["Lag1"])
```

```
# 方差  
np.var(data_set)
```

```
# 标准差  
np.std(data_set)
```

```
# 极差  
np.ptp(data_set["Lag1"])
```

```
# 分位数  
q1=data_set.quantile(0.25)  
q2=data_set.quantile(0.5)  
q3=data_set.quantile(0.75)
```

```
# 汇总统计  
data_set.describe()
```





描述性统计

- 描述性统计关注的是什么？

- 中心位置：均值、中位数、众数
- 发散程度：极差、方差、标准差、变异系数(=STD/mean)
- 偏差程度：Z分数

$$Z\text{-Score} = \frac{X - \text{Mean}}{\text{STD}}$$

Z分数可以用在
什么场景？

```
from numpy import mean, ptp, var, std
```

```
#极差  
ptp(data)  
#方差  
var(data)  
#标准差  
std(data)  
#变异系数  
std(data) / mean(data)
```

```
from numpy import mean, std
```

```
#计算第一个值的z-分数  
(data[0]-mean(data)) / std(data)
```





描述性统计也可以画成图

- 箱形图（盒式图、盒须图）
- 小提琴图

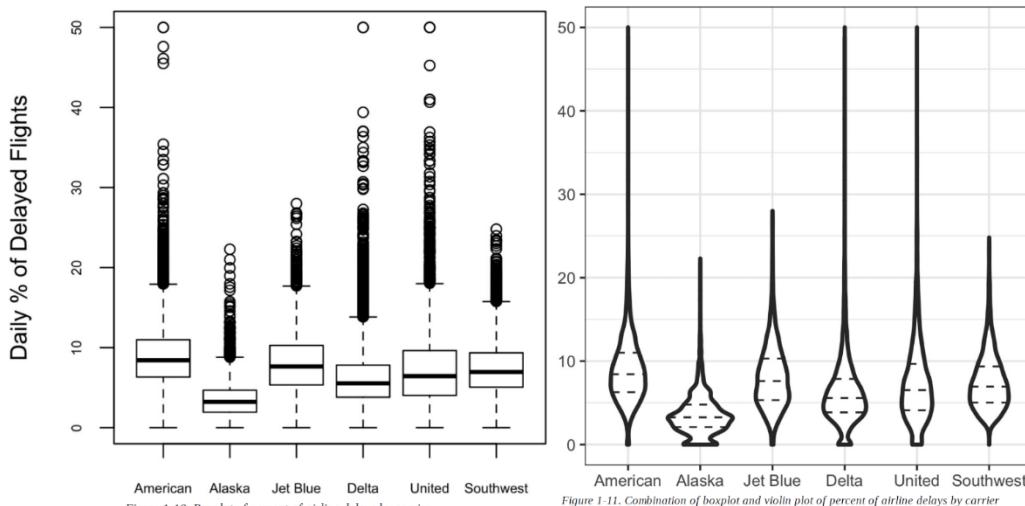
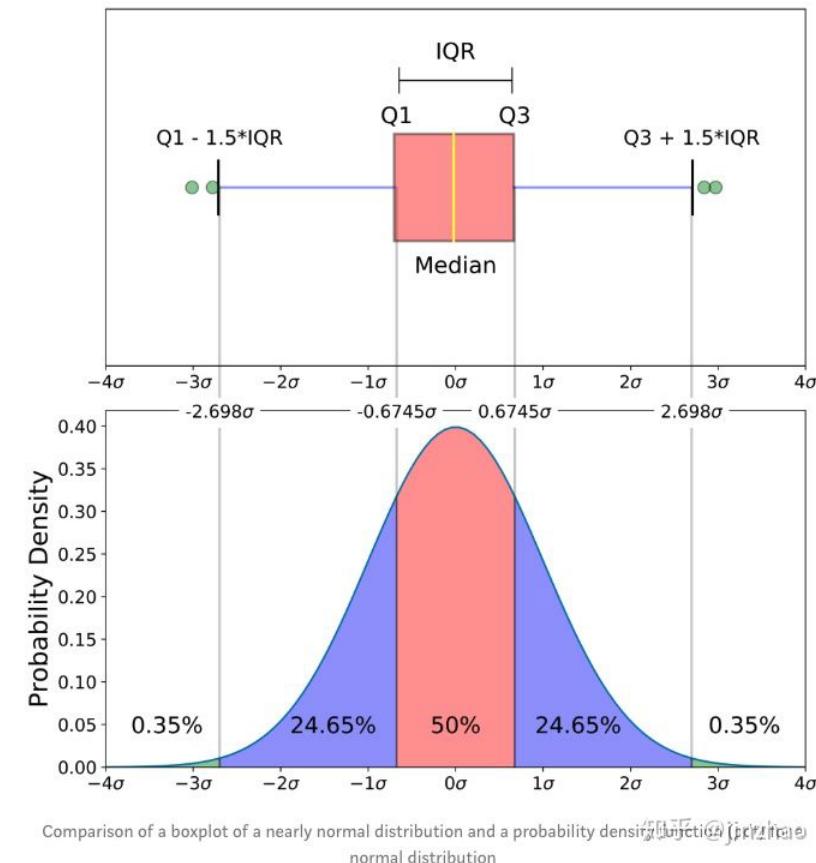


Figure 1-10. Boxplot of percent of airline delays by carrier

Figure 1-11. Combination of boxplot and violin plot of percent of airline delays by carrier

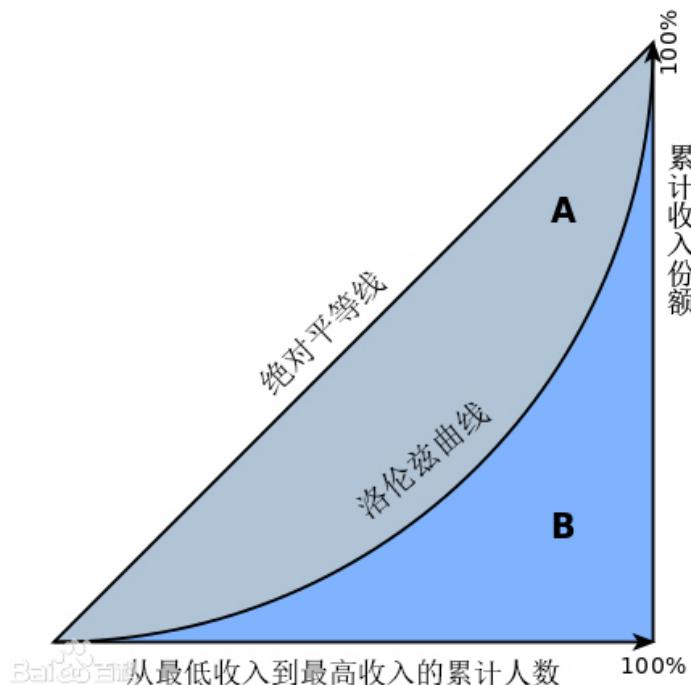




描述性统计还可以做其他的分析

- 例如：均衡性
 - 基尼系数（Gini index）

$$Gini = \frac{S_A}{S_A + S_B}$$



在宏观经济领域：

- 基尼系数最大为“1”，最小等于“0”。基尼系数越接近0表明收入分配越是趋向平等。国际惯例把0.2以下视为收入绝对平均，0.2-0.3视为收入比较平均；0.3-0.4视为收入相对合理；0.4-0.5视为收入差距较大，当基尼系数达到0.5以上时，则表示收入悬殊。



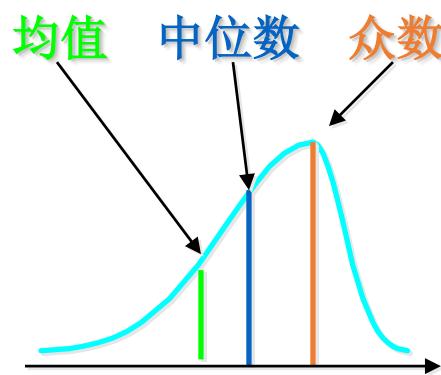
一个变量的分析

- 描述性统计：最值、均值、百分位数、众数
- 分布
 - 表格或柱状图
 - 直方图
 - 密度分布
 - 累积分布
 - 顺便一提：正态分布如何检验？
- 对于同一个变量的多组样本
 - 参数检验
 - 非参数检验

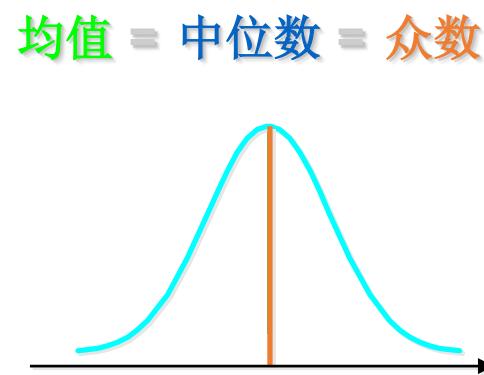




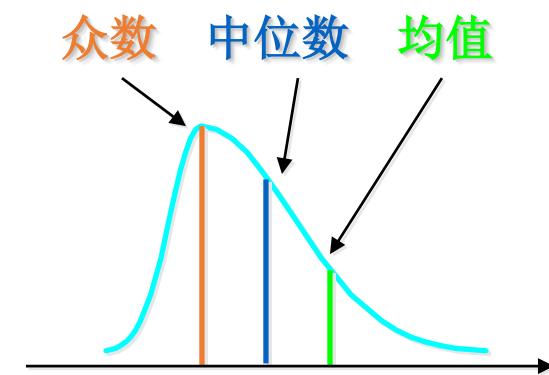
分布



左偏分布



对称分布

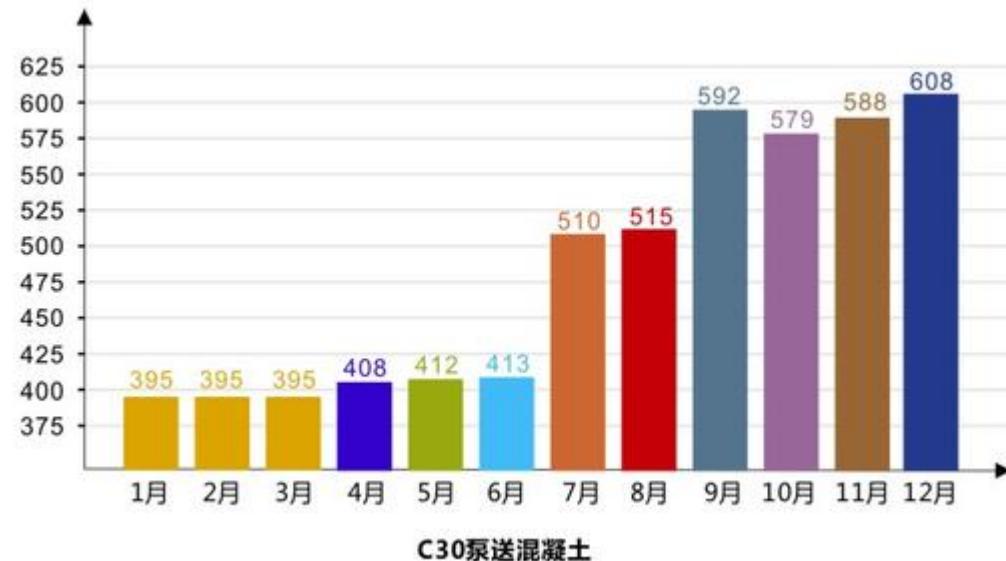


右偏分布

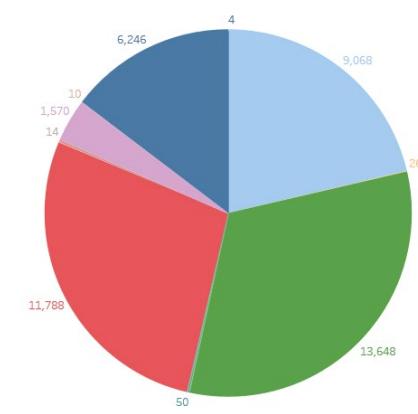


分布

- 表格（近似等效于柱状图）
- 直方图
- 密度分布
- 累积分布



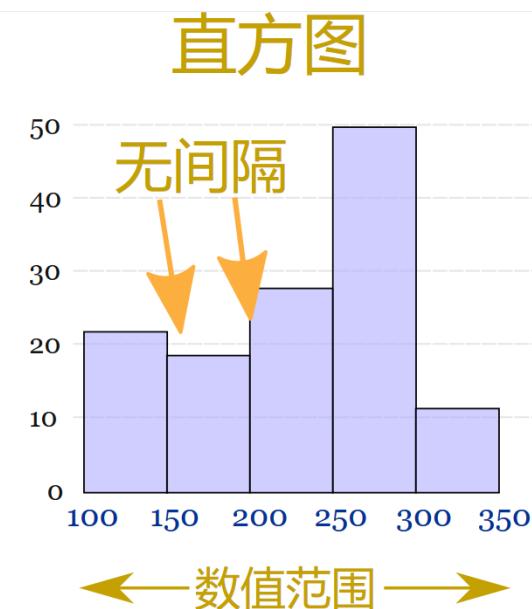
Number of authors	Proportion	Number of authors	Proportion
1	10,664,197 (16.77%)	7	3,014,209 (4.74%)
2	12,495,615 (19.65%)	8	1,984,037 (3.12%)
3	11,370,056 (17.88%)	9	1,252,741 (1.97%)
4	9,068,065 (14.26%)	10	826,682 (1.30%)
5	6,588,019 (10.36%)	11	508,727 (0.80%)
6	4,661,214 (7.33%)	>11	1,157,356 (1.82%)





分布

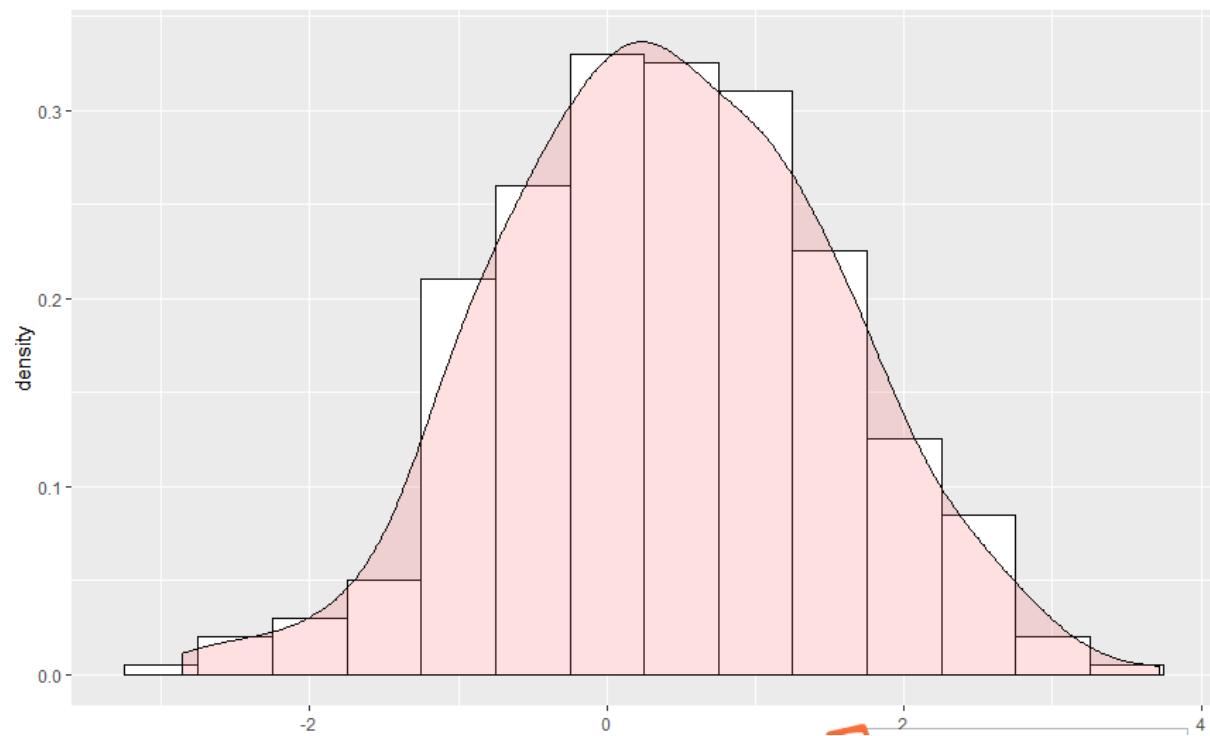
- 表格、柱状图
- 直方图
- 密度分布
- 累积分布





分布

- 表格、条形图
- 直方图
- 密度分布
- 累积分布

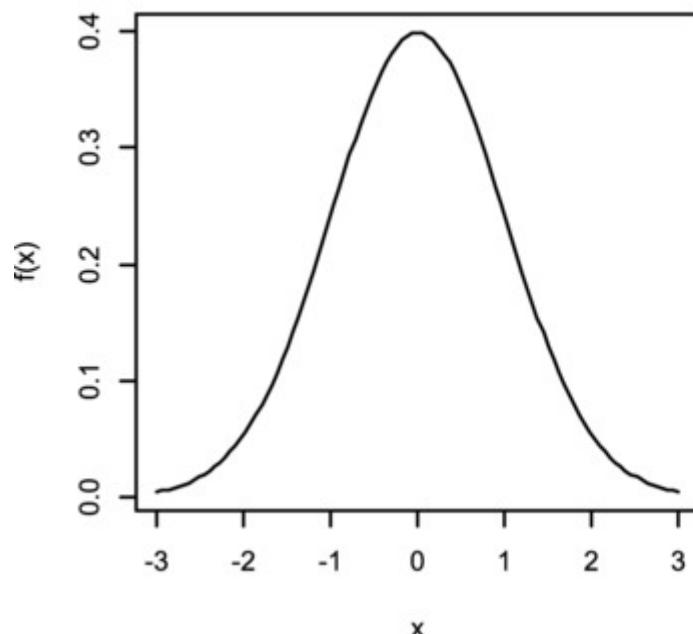




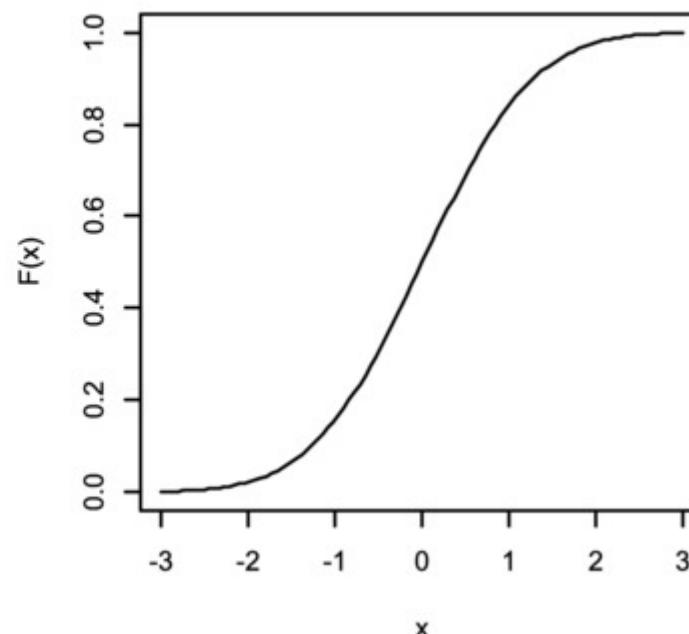
分布

- 表格、条形图
- 直方图
- 密度分布 (PDF)
- 累积分布 (CDF) 或互补累积分布 (CCDF)

Probability density function



Cumulative distribution function

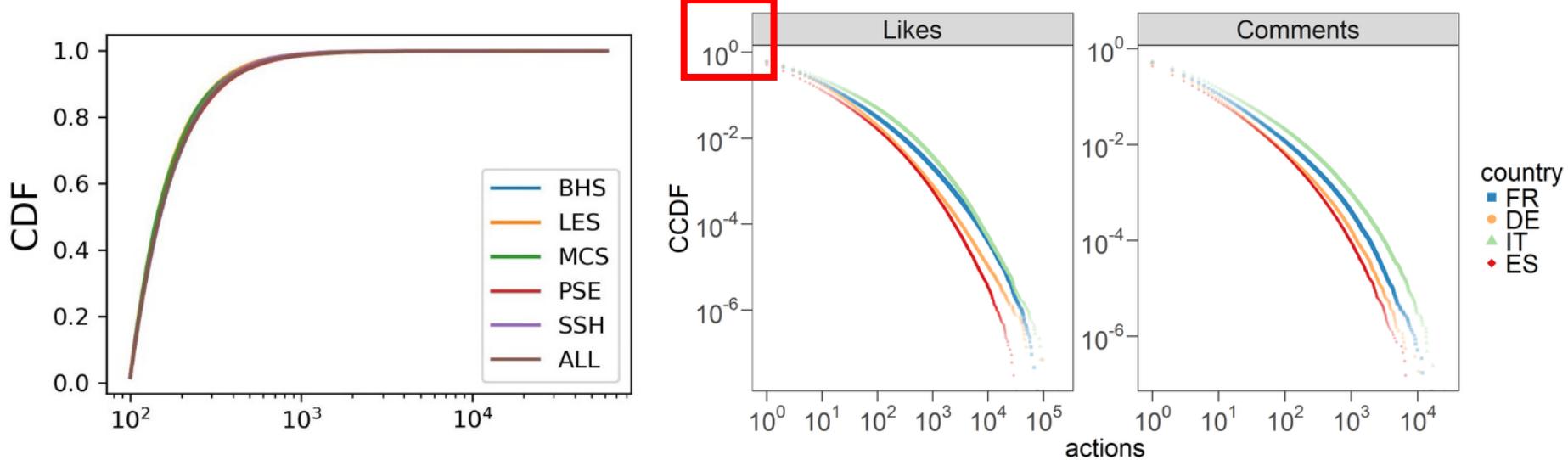




分布

- 表格、条形图
- 直方图
- 密度分布 (PDF)
- 累积分布 (CDF) 或互补累积分布 (CCDF, Complementary Cumulative Distribution Function)

BHS: 生物和健康科学
LES: 生命和地球科学
MCS: 数学和计算机科学
PSE: 物理科学和工程
SSH: 人文社会科学
ALL: 全部





正态分布的检验

- 数值法
- 图示法
 - 直方图法
 - PP图法 (Percentile-percentile plot)/QQ图法 (Quantile-quantile plot)
- 统计检验法
 - Shapiro-Wilk检验
 - Anderson-Darling检验
 -





正态分布的检验：数值法

偏度 (Skewness) 可以用来度量随机变量概率分布的不对称性。

公式：

$$S = \frac{1}{n} \sum_{i=1}^n \left[\left(\frac{X_i - \mu}{\sigma} \right)^3 \right]$$

其中 μ 是均值， σ 是标准差。

计算例子：

一组数据为1、2、2、4、1，均值为2，标准差约为1.22，所以偏度为

$$\begin{aligned} S &= \frac{1}{5} \times \left[\left(\frac{1-2}{1.22} \right)^3 + \left(\frac{2-2}{1.22} \right)^3 + \dots + \left(\frac{4-2}{1.22} \right)^3 \right] \\ &\approx 1.36 \end{aligned}$$

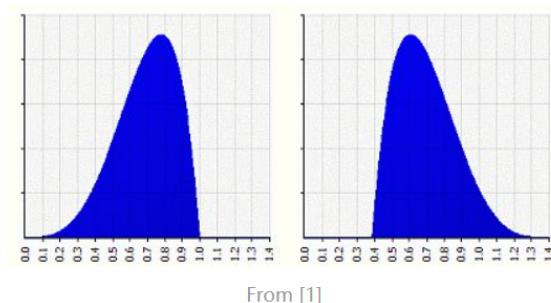
几何意义：

偏度的取值范围为 $(-\infty, +\infty)$

当偏度 <0 时，概率分布图左偏。

当偏度 $=0$ 时，表示数据相对均匀的分布在平均值两侧，不一定是绝对的对称分布。

当偏度 >0 时，概率分布图右偏。



例如上图中，两个概率分布图都是均值=0.6923，标准差=0.1685的，但是他们的形状是不一样的，左图偏度=-0.537，形状左偏，右图偏度=0.537，形状右偏。





正态分布的检验：数值法

峰度 (Kurtosis) 可以用来度量随机变量概率分布的陡峭程度。

公式：

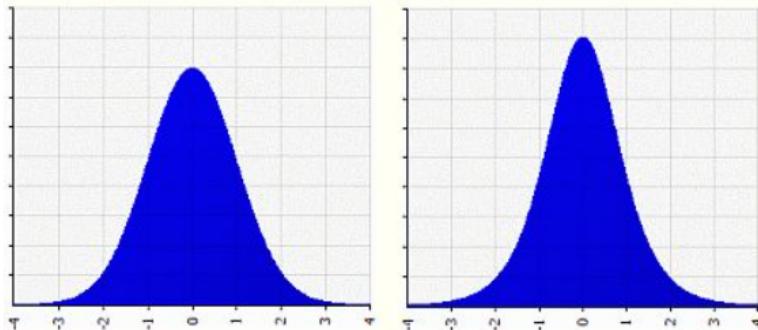
$$K = \frac{1}{n} \sum_{i=1}^n \left[\left(\frac{X_i - \mu}{\sigma} \right)^4 \right]$$

其中 μ 是均值， σ 是标准差。

几何意义：

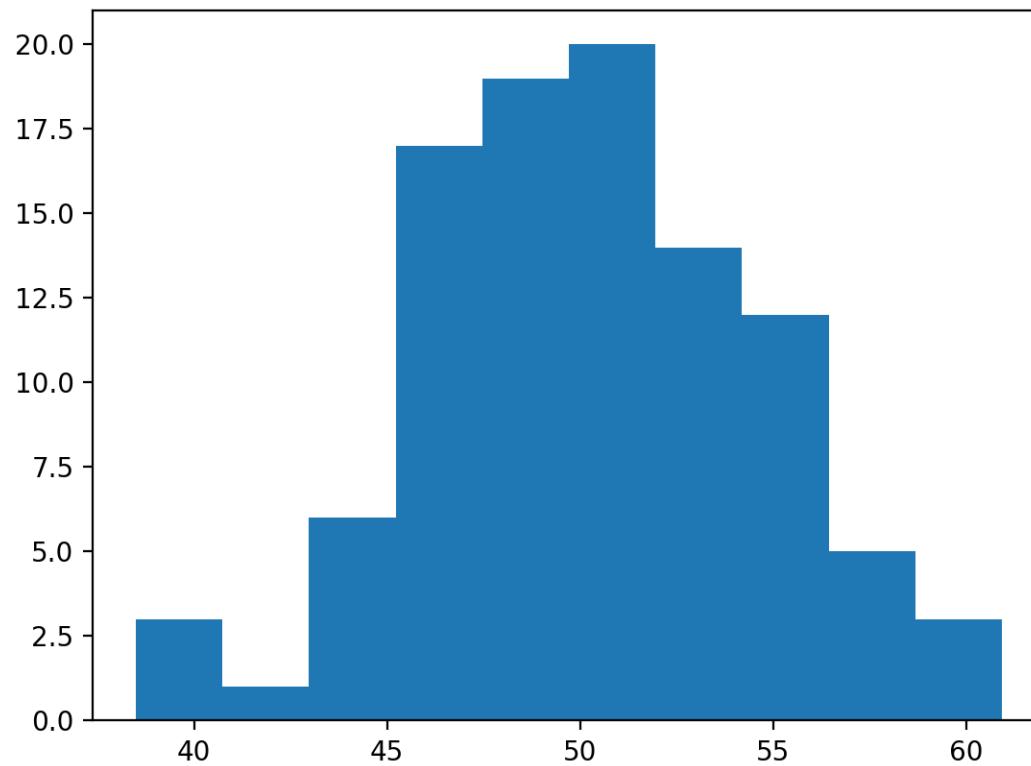
峰度的取值范围为 $[1, +\infty)$ ，完全服从正态分布的数据的峰度值为 3，峰度值越大，概率分布图越高尖，峰度值越小，越矮胖。

- 经验上，正态分布要求：
 - 峰度绝对值 ≤ 10 ，偏度绝对值 ≤ 3





正态分布的检验：直方图法





正态分布的检验：PP图/QQ图

- 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，用于直观地检测样本数据是否符合某一概率分布。
- Q-Q图：P-P图是用分布的累计进行比较，而Q-Q图用的是分布的分位数来做检验。





正态分布的检验：QQ图

- 用途

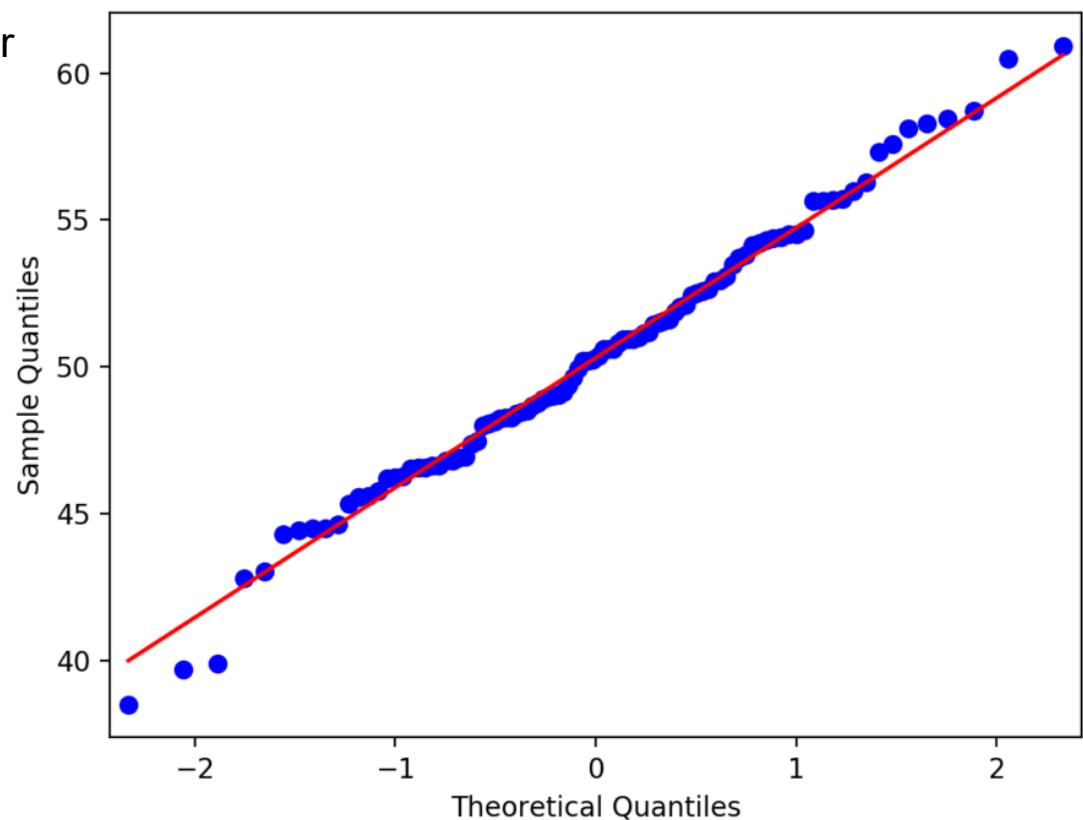
- 检验一组数据是否服从某一分布
- 检验两个分布是否服从同一分布





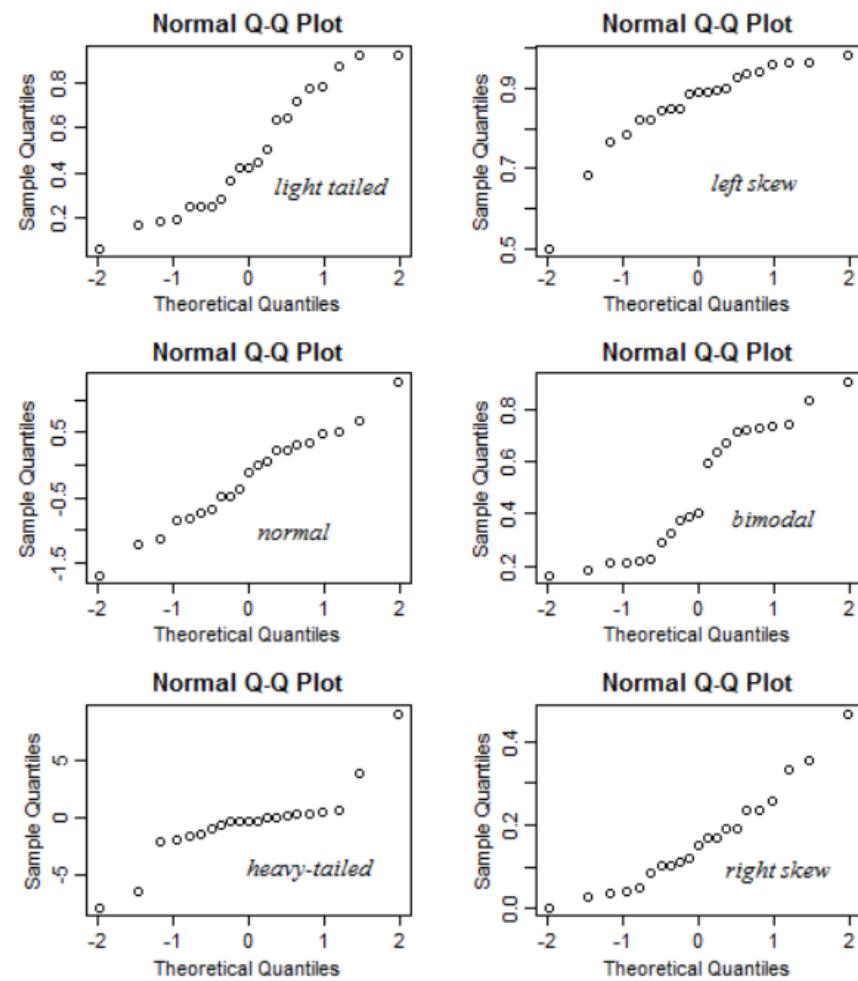
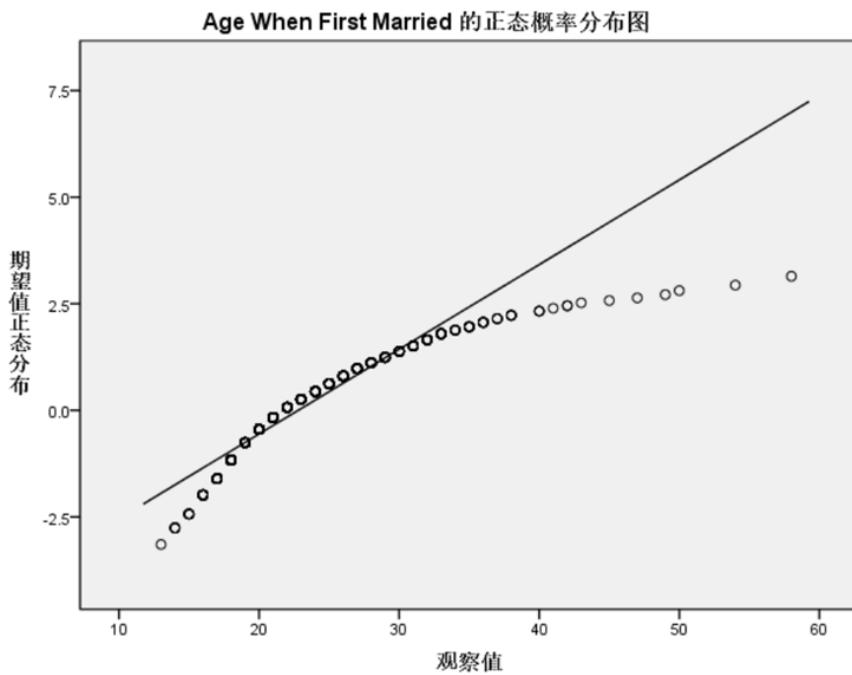
正态分布的检验：QQ图

```
from numpy.random import seed  
from numpy.random import randn  
from statsmodels.graphics.gofplots import qqplot  
from matplotlib import pyplot  
# seed the random number generator  
seed(1)  
# generate univariate observations  
data = 5 * randn(100) + 50  
# q-q plot  
qqplot(data, line='s')  
pyplot.show()
```





正态分布的检验：QQ 图



QQ plots drawn from different types of data which are not quite normal



正态分布的检验：Shapiro-Wilk检验

- 一般用于小样本 ($n \leq 50$)
- 假设：该变量是正态分布的
 - 能够拒绝这个假设呢？来做统计检验
- 返回两个量：
 - SW统计值：和阈值相比即可（阈值通过查表得到）
 - p值：如果小于某值（如0.01、0.05、0.10等），则原数据不正态分布（拒绝原假设）；主要看p值





正态分布的检验：Shapiro-Wilk检验

```
# Shapiro-Wilk Test
from numpy.random import seed
from numpy.random import randn
from scipy.stats import shapiro
# seed the random number generator
seed(1)
# generate univariate observations
data = 5 * randn(100) + 50
# normality test
stat, p = shapiro(data)
print('Statistics=%.3f, p=%.3f' % (stat, p))
# interpret
alpha = 0.05
if p > alpha:
    print('Sample looks Gaussian (fail to reject H0)')
else:
    print('Sample does not look Gaussian (reject H0)')
```

结果：

Statistics=0.992, p=0.822
Sample looks Gaussian (fail to reject H0)





正态分布的检验：Anderson-Darling检验

- 不仅可以看某分布是否正态，还可以看某分布是否是指数、逻辑斯蒂分布...
 - `dist='norm'`
 - `'norm', 'expon', 'logistic'.....`





正态分布的检验：Anderson-Darling检验

```
# Anderson-Darling Test
from numpy.random import seed
from numpy.random import randn
from scipy.stats import anderson
# seed the random number generator
seed(1)
# generate univariate observations
data = 5 * randn(100) + 50
# normality test
result = anderson(data)
print('Statistic: %.3f' % result.statistic)
p = 0
for i in range(len(result.critical_values)):
    sl, cv = result.significance_level[i], result.critical_values[i]
    if result.statistic < result.critical_values[i]:
        print('%.3f: %.3f, data looks normal (fail to reject H0)' % (sl, cv))
    else:
        print('%.3f: %.3f, data does not look normal (reject H0)' % (sl, cv))
```

结果：

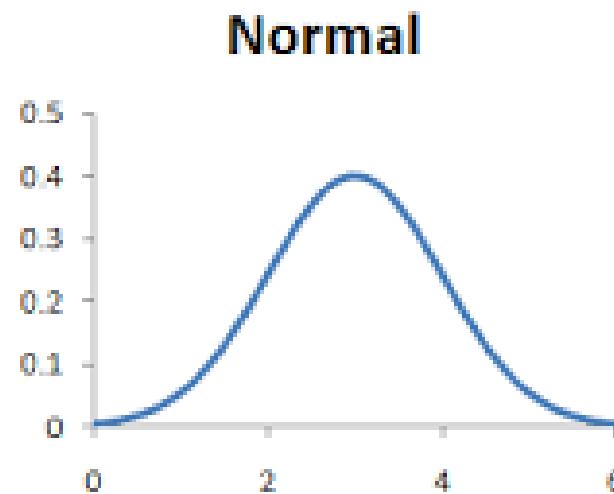
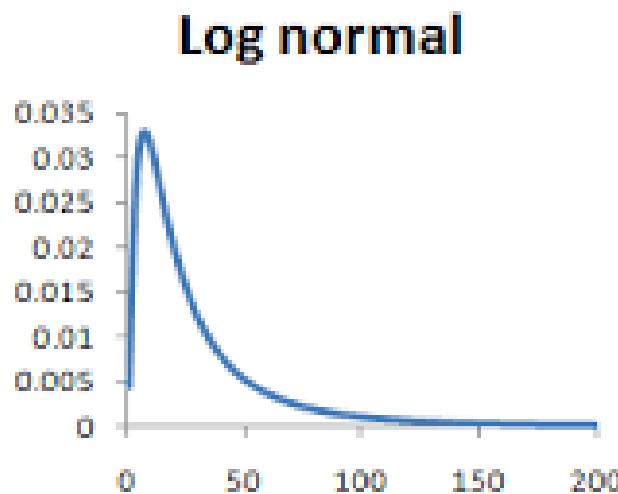
Statistic: 0.220
15.000: 0.555, data looks normal (fail to reject H0)
10.000: 0.632, data looks normal (fail to reject H0)
5.000: 0.759, data looks normal (fail to reject H0)
2.500: 0.885, data looks normal (fail to reject H0)
1.000: 1.053, data looks normal (fail to reject H0)





非正态转化为正态分布

- 可以尝试将非正态数据转化为正态数据，如取对数、开根号等





一个变量的分析

- 描述性统计：最值、均值、百分位数、众数
- 分布
 - 表格或柱状图
 - 直方图
 - 密度分布
 - 累积分布
 - 顺便一提：正态分布如何检验？
- 对于同一个变量的多组样本
 - 参数检验
 - 非参数检验





参数检验和非参数检验

- 如果某变量是（近似）正态分布的，一般说来可以使用参数检验（如t检验），否则一般使用非参数检验。
 - 参数检验：**假定随机样本来自某种已知分布的总体，并对总体分布的参数作检验
 - 例如：t检验、F检验
 - 非参数检验：**对总体分布不作严格规定
 - 例如：卡方检验、K-S检验

比较项目	参数检验	非参数检验
检验对象	总体参数	总体分布和参数
总体分布	正态分布	未知
数据类型	连续数据	连续数据或离散数据
检验效能	较高	较低





参数检验和非参数检验

- 参数/非参数检验可以针对多个变量，目前先只关注一个变量！



t检验

- 什么是t检验：
 - t检验是一种适合小样本 ($n \leq 30$) 的统计分析方法，通过比较不同数据的均值，研究两组数据的均值是否存在差异
- t检验的用途：
 - 单样本t检验：用于比较一组数据与一个特定数值之间的差异情况
 - 配对样本t检验：用于检验有一定对应关系的样本之间的差异情况，需要两个样本数相等
 - 独立样本t检验：两个样本数可以不等



单样本t检验

- 示例：
 - 已知某工厂生产的一种点火器平均寿命大于1200次为合格产品，现在质检部随机抽取了20个点火器进行试验，结果寿命分别为（单位：次）：
 - 809, 1250, 689, 1541, 995, 1234, 1024, 920, 777, 2510, 2301, 540, 850, 895, 1024, 1000, 1025, 863, 875, 1105
- 单样本t检验用**stats.ttest_1samp**函数

```
from scipy import stats
import numpy as np
#单样本t检验
sample=[809, 1250, 689, 1541, 995, 1234, 1024, 920, 777, 2510,
2301, 540, 850, 895, 1024, 1000, 1025, 863, 875, 1105]
sample = np.asarray(sample)
m = np.mean(sample)
print("样本均值:",m)
r = stats.ttest_1samp(sample, 1200, axis=0)
print("statistic:", r.__getattribute__ ("statistic"))
print("pvalue:", r.__getattribute__ ("pvalue"))
```

结果：

样本均值: 1111.35
statistic: -0.8043067483882222
pvalue: 0.4311691484589055





配对样本t检验

- 是什么
 - 用于检验有一定**对应关系**的样本之间的差异情况，需**要两组样本数相等**
- 示例：
 - 现有2种血压计，为研究其测量性能是否有显著差异，分别使用这两种血压计测量15个人的血压值：

受测人编号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
血压计A	68	85	123	74	88	128	63	115	110	93	99	75	89	68	140
血压计B	60	88	132	70	95	115	74	115	121	86	96	71	96	70	143





配对样本t检验

- 使用stats.ttest_rel进行成对数据的t检验

```
from scipy import stats  
import numpy as np
```

#成对样本t检验

```
sample1=[68, 85, 123, 74, 88, 128, 63, 115, 110, 93, 99, 75, 89, 68, 140]  
sample2=[60, 88, 132, 70, 95, 115, 74, 115, 121, 86, 96, 71, 96, 70, 143]  
sample1 = np.asarray(sample1)  
sample2 = np.asarray(sample2)  
r = stats.ttest_rel(sample1,sample2)  
print("statistic:", r.__getattribute__("statistic"))  
print("pvalue:", r.__getattribute__("pvalue"))
```

Statistic<0，说明A血压计结果比B略低
P很大：没有显著差异

结果：

```
statistic: -0.492910604649229  
pvalue: 0.6297167086230713
```





配对样本t检验

- 常见的使用场景有：
 - 同一对象**处理前后的对比**（同一组人员采用同一种减肥方法前后的效果对比）
 - 同一对象**采用两种方法的结果的对比**（同一组人员分别服用两种减肥药后的效果对比）
 - **配对**的两个对象分别接受两种处理后的结果对比（两组人员，按照体重进行配对，服用不同的减肥药，对比服药后的两组人员的体重）



独立样本t检验

- 示例：

- 现市场上有2种蓄电池，为研究哪种蓄电池比较好，分别抽取两种蓄电池若干件，测试其续航时间（单位：h），根据抽样结果分析两种蓄电池续航是否有显著差异

A型蓄电池	5.5	5.6	6.3	4.6	5.3	5.0	6.2	5.8	5.1	5.2	5.9
B型蓄电池	3.8	4.3	4.2	4.0	4.9	4.5	5.2	4.8	4.5	3.9	3.7





独立样本t检验

- 独立2个样本t检验用**stats.ttest_ind**函数

```
from scipy import stats  
import numpy as np
```

```
#独立2个样本t检验  
sample1=[5.5, 5.6, 6.3, 4.6, 5.3, 5.0, 6.2, 5.8, 5.1, 5.2, 5.9]  
sample2=[3.8, 4.3, 4.2, 4.0, 4.9, 4.5, 5.2, 4.8, 4.5, 3.9, 3.7, 4.6]  
sample1 = np.asarray(sample1)  
sample2 = np.asarray(sample2)  
r = stats.ttest_ind(sample1, sample2)  
print("statistic:", r.__getattribute__("statistic"))  
print("pvalue:", r.__getattribute__("pvalue"))
```

结果:

```
statistic: 5.484377451921326  
pvalue: 1.9279192737974777e-05
```

Statistic>0，说明A型的均值大于B型的均值；
P很小：有显著差异





F检验 (ANOVA, 方差检验)

- t检验关注两组之间的某一变量（的均值）是否有显著差异
- 方差分析就是对试验数据进行分析，检验方差相等的多个正态总体均值是否相等，进而判断各因素对试验指标的影响是否显著。
 - 其原理认为不同处理组的均值间的差别基本来源有两个：实验条件和随机误差。其思想为通过分析研究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。





F检验 (ANOVA, 方差检验)

- 单因素方差分析
 - 适用于问卷数据和实验数据，实验中只有一个因素改变的样本。判断该因素对样本的影响因素是否显著。
- 双因素方差分析
 - 适用于实验数据，实验中有两个因素改变的样本。
- 多因素方差分析
 - 适用于实验数据，实验中有多个因素改变的样本。





使用Python进行单因素F检验

- **示例：**下面以射击比赛为例，三位选手分别成绩如下：
 - Pat - 5, 4, 4, 3, 9, 4
 - Jack - 4, 8, 7, 5, 1, 5
 - Alex - 9, 9, 8, 10, 4, 10
- 基于上述数据，我们希望判断上述三个选手中成绩最好的。原假设：三个选手的成绩无显著差异。

$$H_0: \mu_1 = \mu_2 = \mu_3$$

- 拒绝原假设的就表示在三个选手中至少有两个人是具有显著差异的。





```
import numpy as np  
from scipy import stats
```

```
data = np.rec.array([  
    ('Pat', 5),  
    ('Pat', 4),  
    ('Pat', 4),  
    ('Pat', 3),  
    ('Pat', 9),  
    ('Pat', 4),  
    ('Jack', 4),  
    ('Jack', 8),  
    ('Jack', 7),  
    ('Jack', 5),  
    ('Jack', 1),  
    ('Jack', 5),  
    ('Alex', 9),  
    ('Alex', 8),  
    ('Alex', 8),  
    ('Alex', 10),  
    ('Alex', 5),  
    ('Alex', 10)], dtype = [('Archer','|U5'),('Score', '<i8')])
```

```
f, p = stats.f_oneway(data[data['Archer'] == 'Pat'].Score,  
                      data[data['Archer'] == 'Jack'].Score,  
                      data[data['Archer'] == 'Alex'].Score)
```

```
print ('One-way ANOVA')  
print ('=====')
```

```
print ('F value:', f)  
print ('P value:', p, '\n')
```

One-way ANOVA

=====

F value: 4.999999999999998

P value: 0.021683749320078414





使用Python进行单因素F检验

- 由于P值 $0.02 < 0.05$;所以我们拒绝原假设，认为三个选择间**至少有两个具有显著性的差异**。虽然知道了三者间具有显著差异，但是我们还是希望能够判断出哪个选手是三人中成绩最好的，**接下来可以采用Tukey's range test进行分析**。



```
from statsmodels.stats.multicomp import pairwise_tukeyhsd  
from statsmodels.stats.multicomp import MultiComparison
```

```
mc = MultiComparison(data['Score'], data['Archer'])  
result = mc.tukeyhsd()
```

结果如下图

```
print(result)
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05						
		group1	group2	meandiff	lower	upper
		reject				
Alex	Jack	-3.3333	-6.5755	-0.0911	True	
Alex	Pat	-3.5	-6.7422	-0.2578	True	
Jack	Pat	-0.1667	-3.4089	3.0755	False	

在上图中，可以看出(Alex&Jack) 和 (Alex&Pat)拒绝原假设，

$$\mu_{Jack} - \mu_{Alex} = -3.3333$$

$$\mu_{Pat} - \mu_{Alex} = -3.5$$

通过均值的差值可以得出结论Alex是三个选择中成绩最好的。





卡方检验 (χ^2 [chi-square] test)

- 是什么
 - 分析**定性（分类）**数据差异性的方法。是一种通过频数进行检验的方法。
- 用途
 - 卡方优度检验（单组样本）：对一列数据进行统计检验，分析单个分类变量**实际观测的比例与期望的比例**是否一致。
 - 交叉表卡方检验（两组独立样本）：研究**两组分类变量的关系**：如性别与看不看直播是否有关系。
 - 配对卡方检验（两组配对样本）：研究实验过程中，用不同方法检测同一批**对象**，看**两个方法的效果是否有显著差异**。





单样本K-S检验

- K-S检验是一种统计检验方法，其通过比较两样的频率分布、或者一个样本的频率分布与特定理论分布（如正态分布）之间的差异大小来推论两个分布**是否来自同一分布**。



单样本K-S检验

```
from scipy import stats  
stats.kstest(rvs, cdf, args=(),...)  
#其中rvs可以是数组、生成数组的函数或者scipy.stats  
里面理论分布的名字  
#cdf可以与rvs一致。若rvs和cdf同是数组，则是比较两数  
组的分布是否一致；一个是数  
组，另一个是理论分布的名字，则是看样本是否否和理论分  
布  
#args是一个元组，当rvs或者cds是理论分布时，这个参数  
用来存储理论分布的参数，如正态分布的mean和std。
```

```
Input: stats.kstest(test, 'norm', args=(test.mean(), test.std()))  
Output: KstestResult(statistic=0.005777479839093713, pvalue=0.8923049615924274)
```





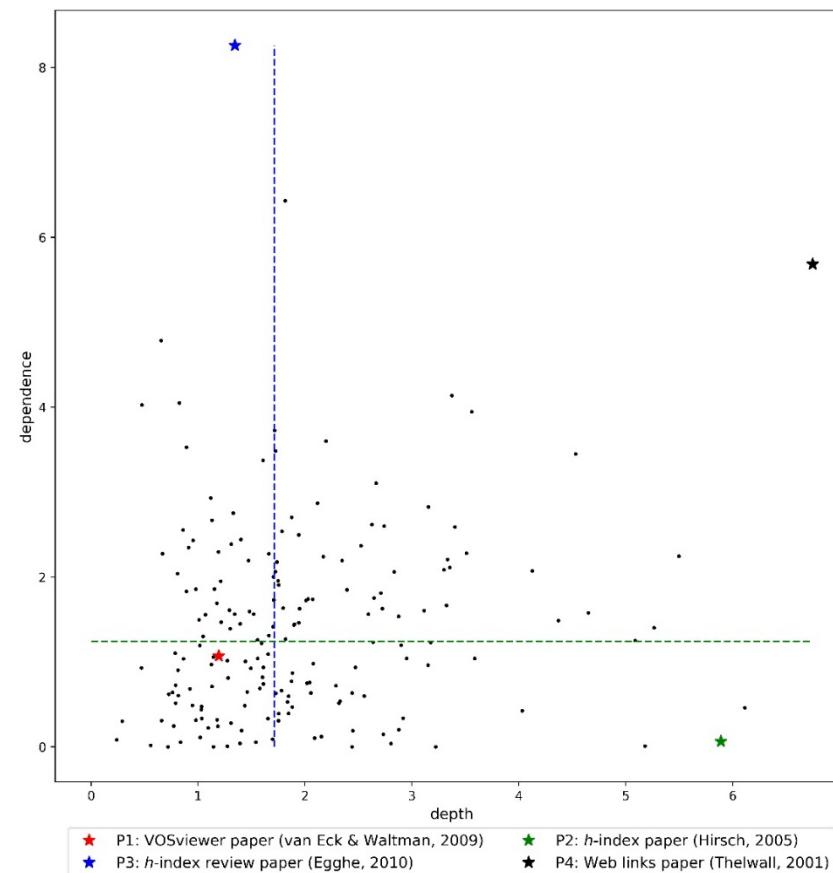
两个变量的分析





两个变量的分析

- 最简单的办法：观察散点图

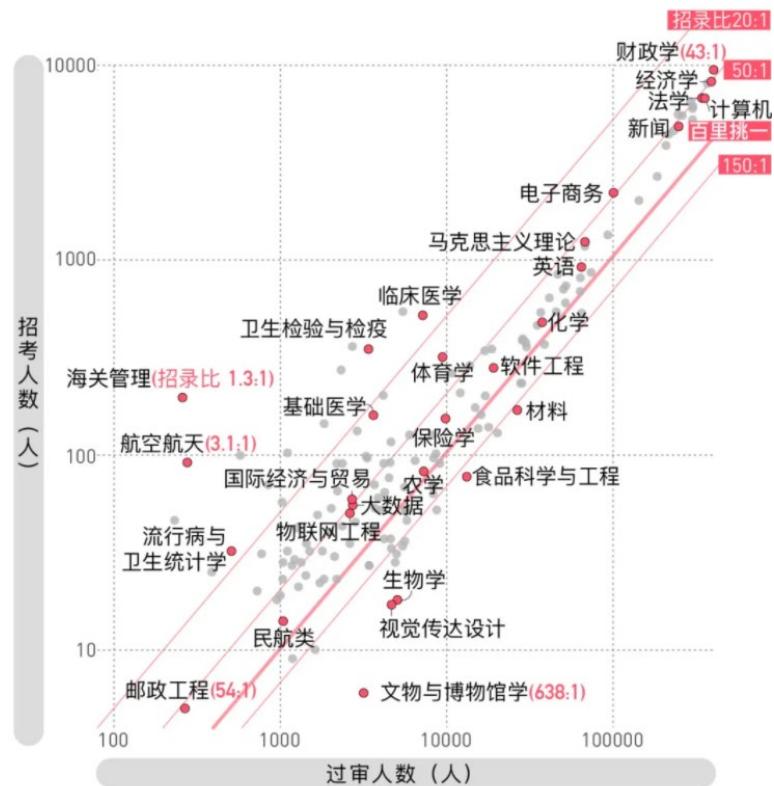




两个变量的分析

与上一页的散点图不同，
这张散点图中的点针对
X和Y轴做了聚合！

- 最简单的办法：观察散点图



注：因符合同一岗位的专业有多个，过审人数存在重复计算。X轴、Y轴为对数坐标。

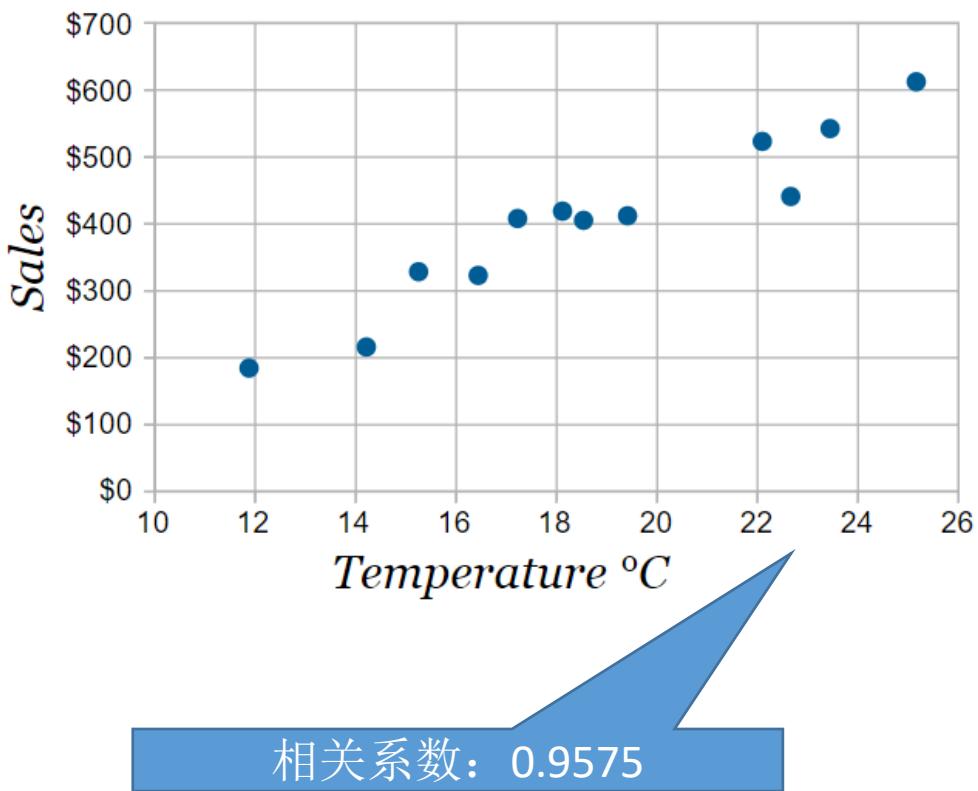
数据来源：中公教育、国家公务员局





两个变量的分析

- 相关系数 (correlation coefficient)



Ice Cream Sales vs Temperature	
Temperature °C	Ice Cream Sales
14.2°	\$215
16.4°	\$325
11.9°	\$185
15.2°	\$332
18.5°	\$406
22.1°	\$522
19.4°	\$412
25.1°	\$614
23.4°	\$544
18.1°	\$421
22.6°	\$445
17.2°	\$408



两个变量的分析

- 相关系数 (correlation coefficient)
 - 皮尔逊Pearson相关系数。要求：连续数据、正态分布、线性关系
 - 斯皮尔曼Spearman相关系数。

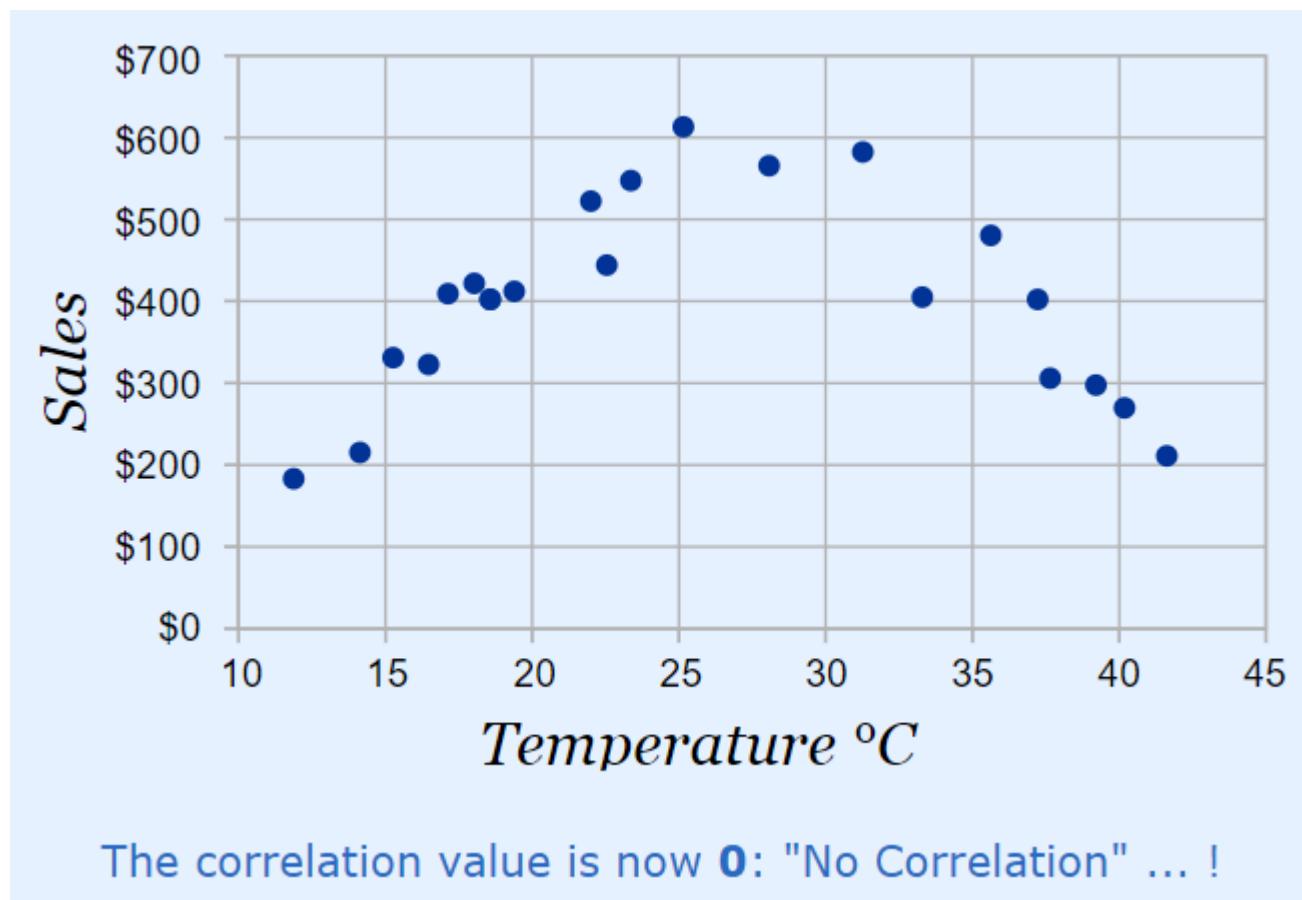
$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$





两个变量的分析

- 相关系数 (correlation coefficient)





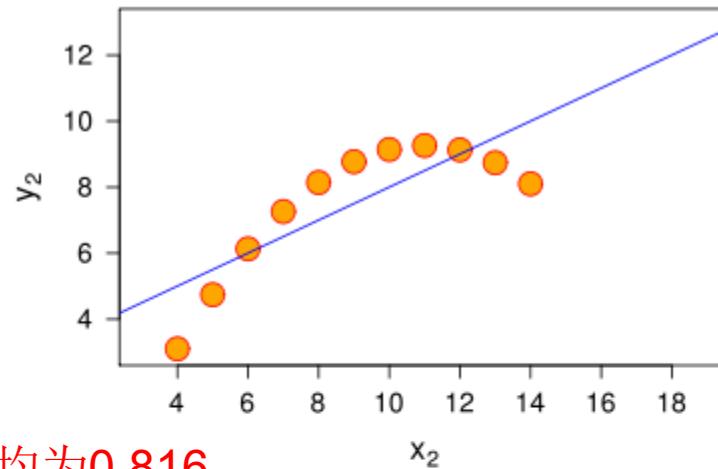
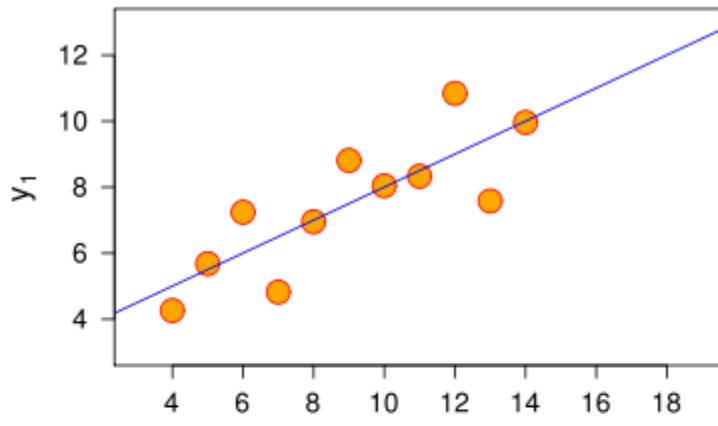
使用Python计算相关系数

- Numpy包
 - `pccs = np.corrcoef(x, y)`
- Scipy包
 - `from scipy.stats import pearsonr`
 - `pccs = pearsonr(x, y)`
- pandas包
 - `DataFrame.corr(method='pearson', min_periods=1)`
 - method : 指定相关系数的计算方式，可选：
`{'pearson', 'kendall', 'spearman'}`
 - `pearson` : 皮尔逊相关系数
 - `kendall` : kendall秩相关系数
 - `spearman` : 斯皮尔曼等级相关系数

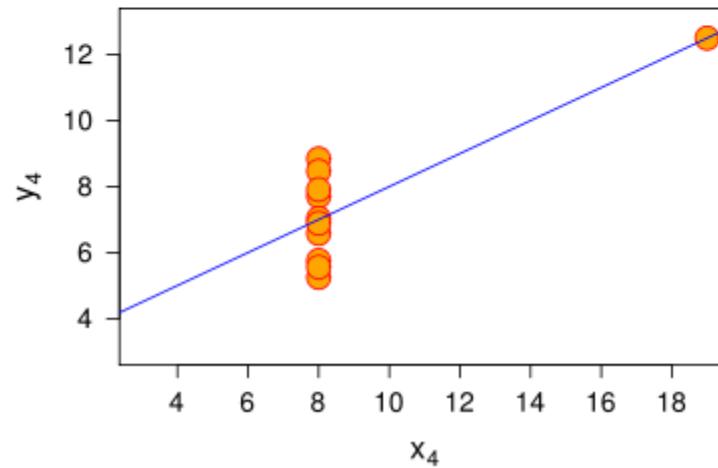
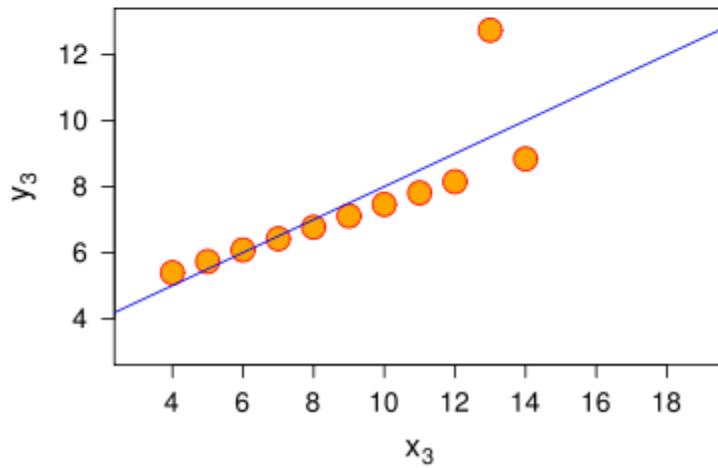




相关系数是万能的吗？



相关系数均为0.816





三个或三个以上变量的分析





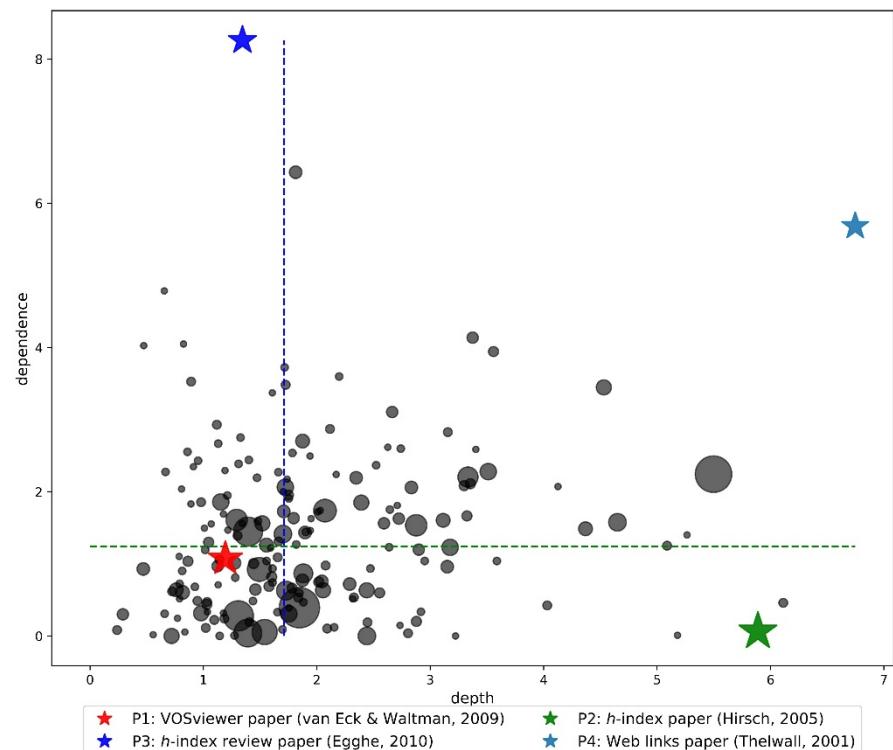
三个或三个以上变量的分析

- 只能用于3个变量：
 - 珍珠图/气泡图、热力图等
- 可以用于3个或以上变量：
 - 两两分别看（相关系数图、散点图矩阵等）
 - 因子分析
 - 聚类分析
 - 判别分析
 - 回归分析



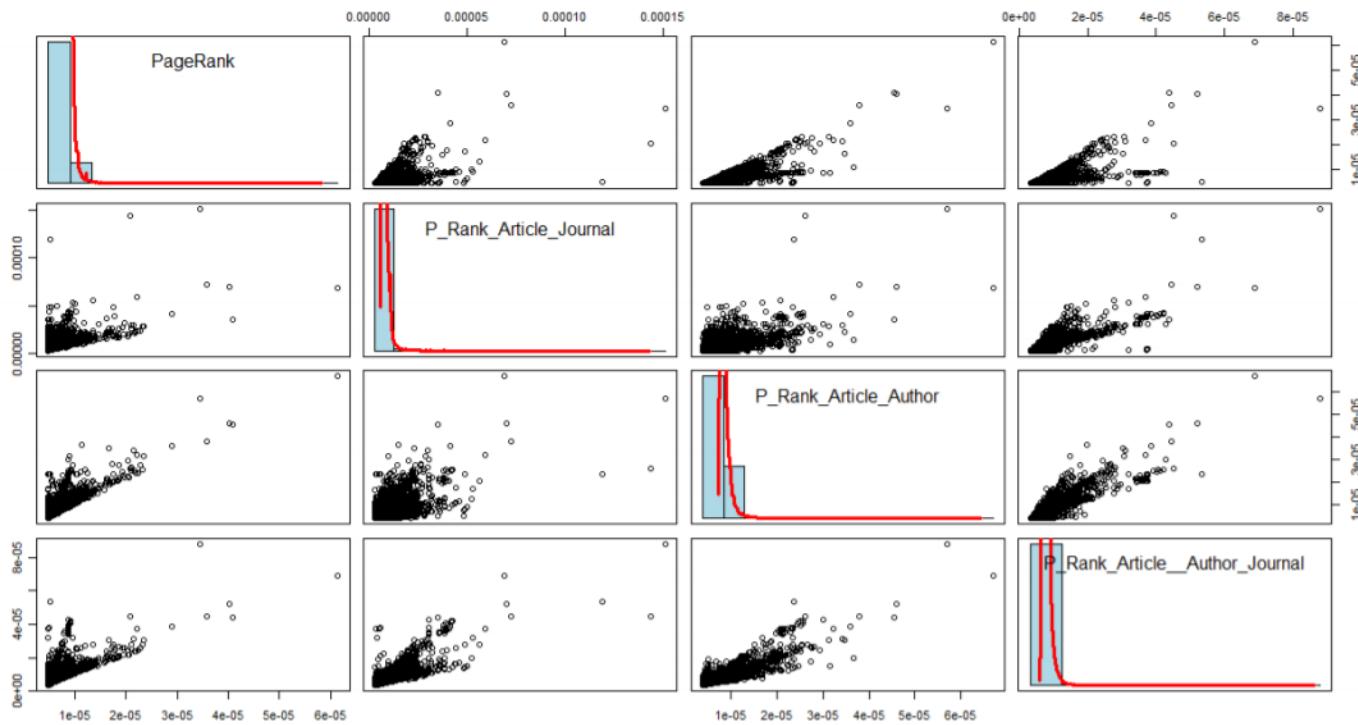
三个变量之间的关系

- 如果还想使用散点图，可能就要进化为珍珠图/气泡图 (bubble plot)
 - 身高、体重、肺活量

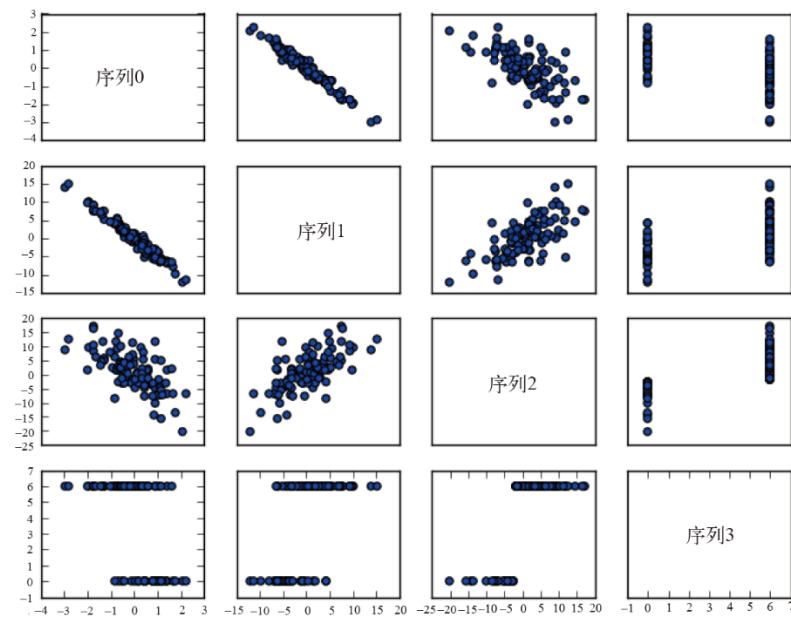




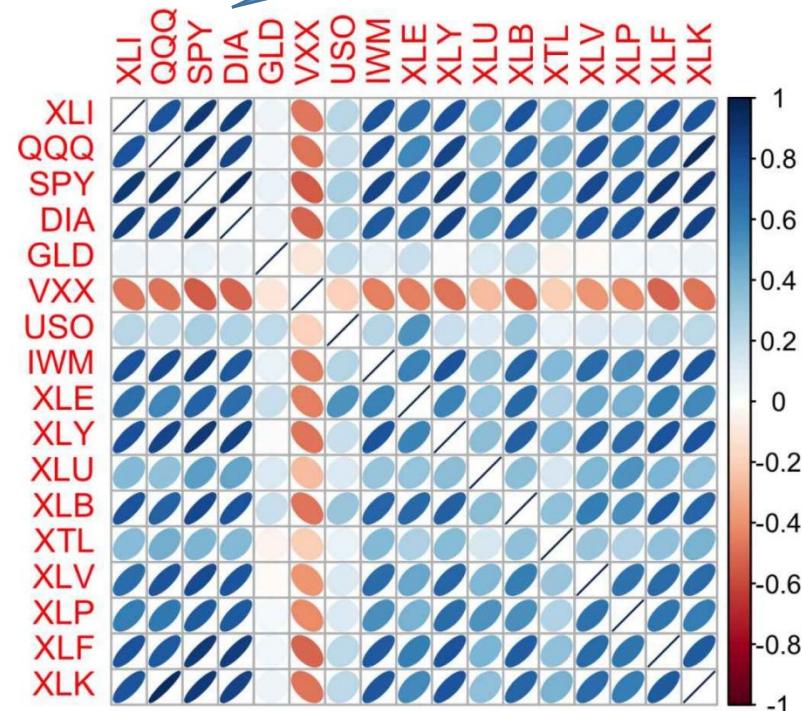
相关系数图/相关矩阵图/热力图/散点图矩阵



相关系数图/相关矩阵图/热力图/散点图矩阵



散点图矩阵



相关系数图



三个或三个以上变量的分析

- 只能用于3个变量：
 - 珍珠图/气泡图、热力图等
- 可以用于3个或以上变量：
 - 两两分别看（相关系数图、散点图矩阵等）
 - 因子分析
 - 聚类分析
 - 判别分析
 - 回归分析



因子分析 (factor analysis)

- 因子分析的目的是浓缩数据
 - 通过研究众多变量之间的内部依赖关系，探求观测数据中的基本结构，并用少数几个虚拟变量来表示基本的数据结构。
 - 这些虚拟变量称为因子 (factor)，能够反映原来众多的观测变量所代表的主要信息，并解释这些观测变量之间的相互依存关系。
 - 因子分析就是研究如何以最少的信息丢失把众多的观测变量浓缩为少数几个因子。





因子分析

- 高中文理科分班的时候，班主任如何推荐学生去文科班/理科班？

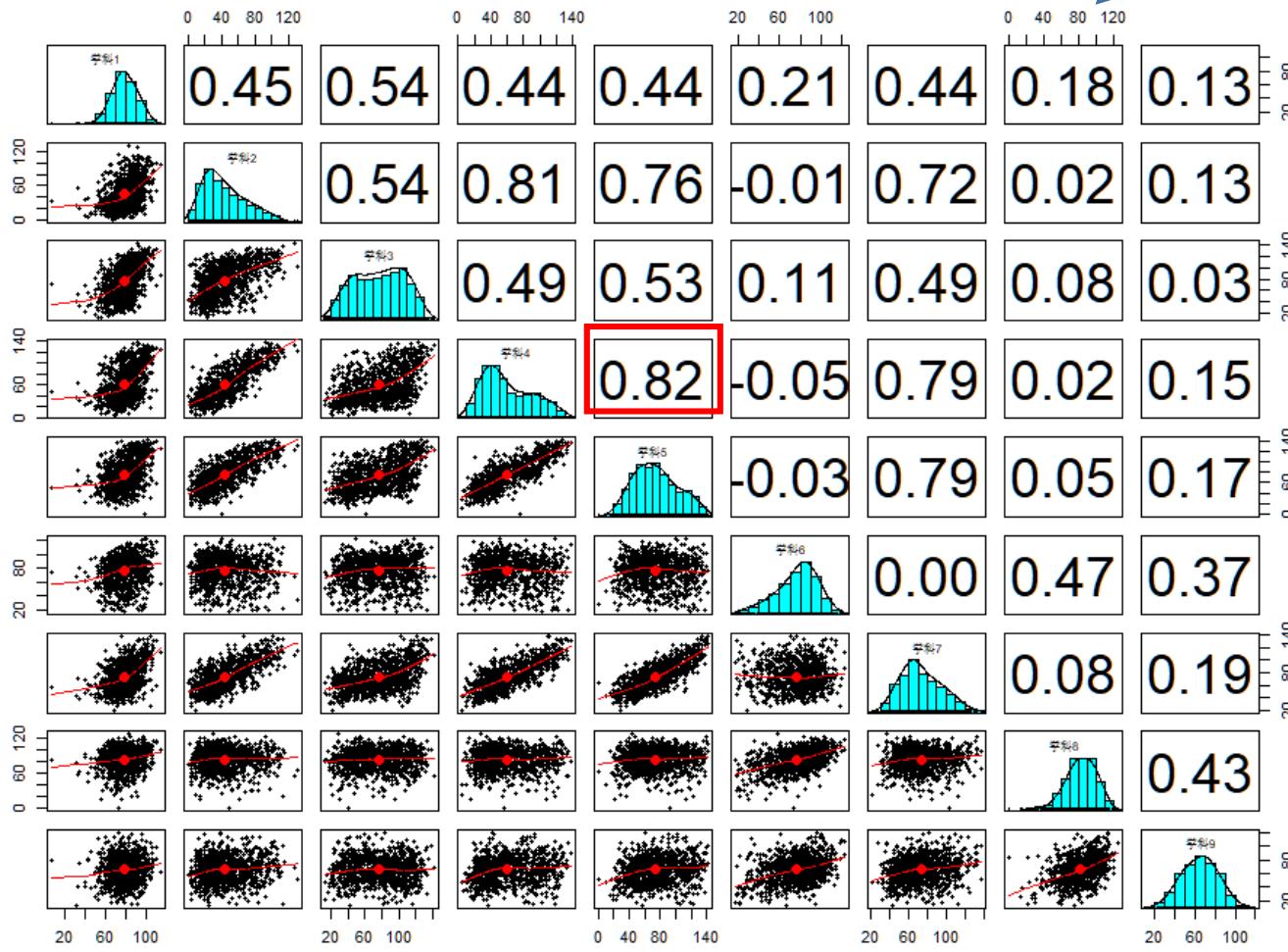
学科1	学科2	学科3	学科4	学科5	学科6	学科7	学科8	学科9
101	109	130.0	130	137	74	133	102	73
97	93	105.5	120	139	83	133	122	90
100	84	129.0	119	122	87	124	89	78
98	89	114.5	127	131	69	126	94	83
101	75	114.0	120	128	88	115	89	88
91	84	126.5	121	130	84	104	100	72
79	83	114.5	119	129	92	122	96	70
94	82	103.0	111	135	89	118	90	76
114	96	87.5	125	125	58	113	100	61
97	78	90.5	132	117	80	115	91	73
102	105	95.5	100	123	59	103	103	81
88	81	94.5	114	135	79	123	85	72





因子分析

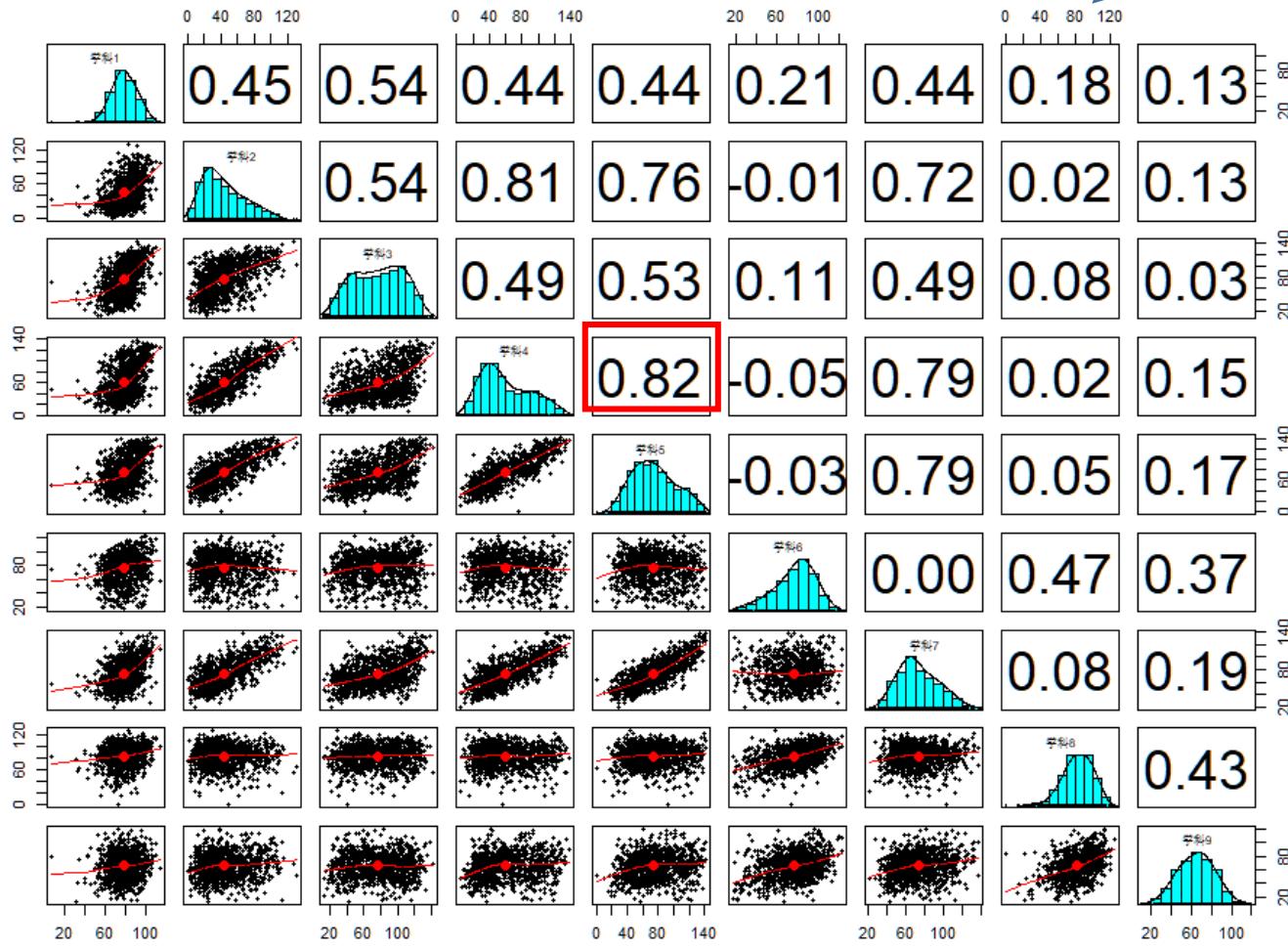
如果变量之间有较强的相关性，我们就可以把它们“打包”到一起作为一个新变量，即“数据降维”





因子分析

如果变量之间有较强的相关性，我们就可以把它们“打包”到一起作为一个新变量，即“数据降维”



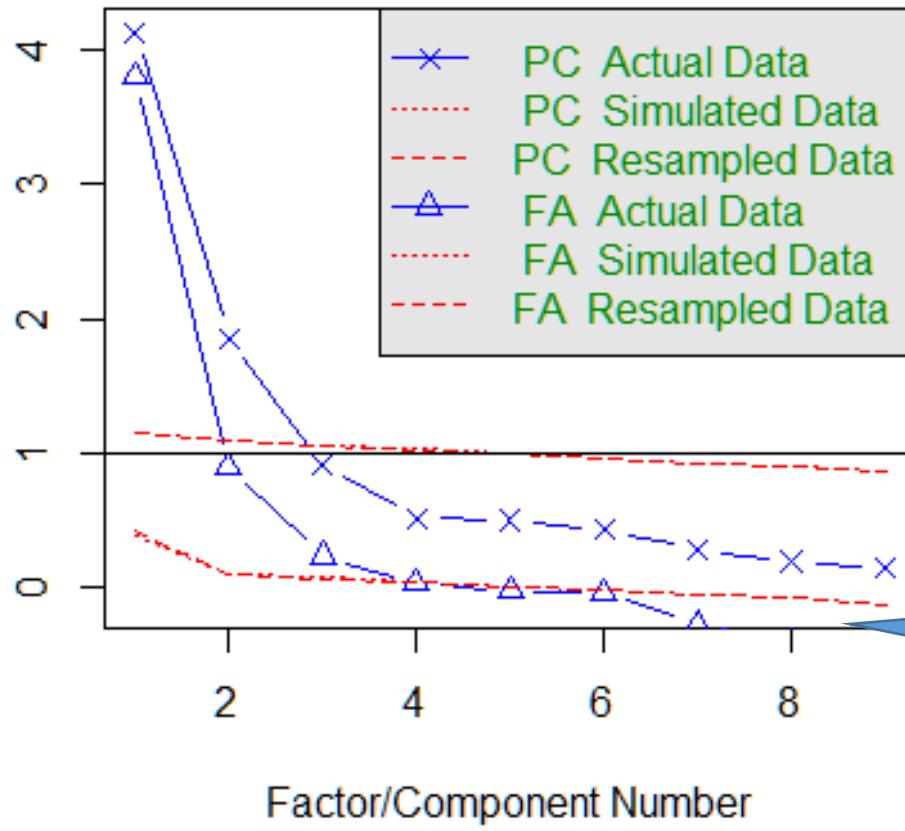
这笔数据可以“打几个包”（几个因子）？



因子分析

eigenvalues of principal components and factor analysis

Parallel Analysis Scree Plots



碎石图

有3个蓝三角在下红线上面





因子分析

ML1 ML2 ML3

语文 0.18 0.12 0.52

数学 0.79 -0.05 0.14

英语 0.12 -0.09 0.79

物理 0.94 -0.03 -0.02

化学 0.85 -0.01 0.08

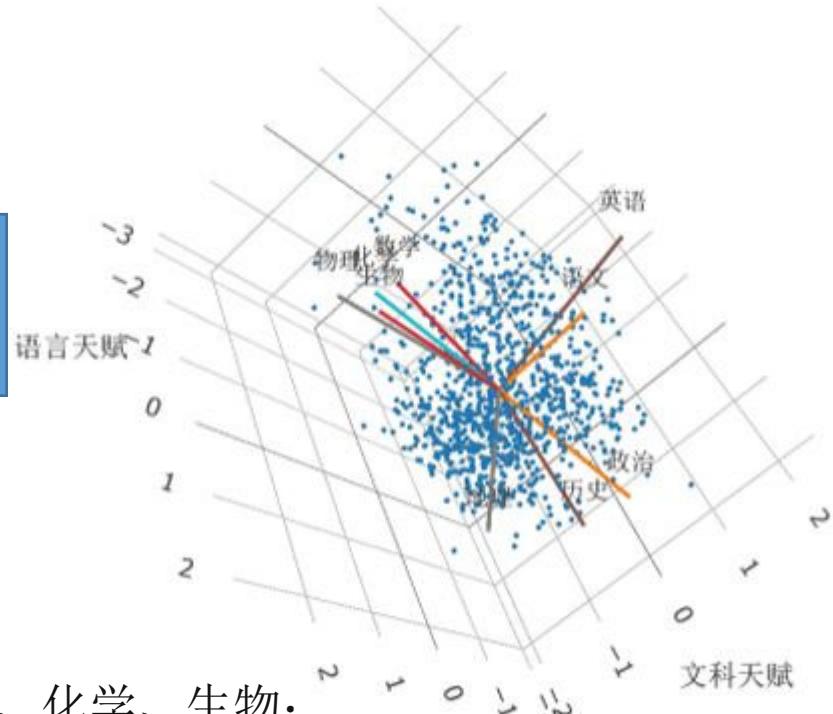
政治 -0.21 0.64 0.19

生物 0.83 0.04 0.05

历史 -0.05 0.70 0.03

地理 0.23 0.64 -0.21

此因子对于这一学科的
“解释力”（负载值）



似乎：

ML1对应数学、物理、化学、生物；

ML2对应政治、历史、地理；

ML3对应语文、英语。

语文	数学	英语	物理	化学	政治	生物	历史	地理	ML1	ML2	ML3
101	109	130.0	130	137	74	133	102	73	2.52143767	0.704916287	1.878961855
97	93	105.5	120	139	83	133	122	90	2.30135028	1.580503818	1.299845376
100	84	129.0	119	122	87	124	89	78	1.98157842	0.720732522	1.736381115
98	89	114.5	127	131	69	126	94	83	2.23887562	0.599880061	1.321480829
101	75	114.0	120	128	88	115	89	88	1.91982889	0.900981004	1.372657204
91	84	126.5	121	130	84	104	100	72	1.88802168	0.689954038	1.544961729
79	83	114.5	119	129	92	122	96	70	1.96969327	0.690534978	1.149036862
94	82	103.0	111	135	89	118	90	76	1.90128857	0.719284152	1.131773723



因子分析的步骤

- 分析KMO和巴特利特（Bartlett）球形检验
 - **分析KMO值**；如果此值高于0.8，则说明非常适合进行因子分析；如果此值介于0.7~0.8之间，则说明比较适合进行因子分析；如果此值介于0.6~0.7，则说明可以进行因子分析；如果此值小于0.6，说明不适合进行因子分析。
 - 如果Bartlett检验对应 p 值小于0.05也说明适合进行因子分析。
- 描述因子提取情况和方差解释率等
 - 特征值（Eigenvalue） >1 的因子一般可以保留。
 - 描述总共提取的因子个数；分析每个因子旋转后的方差解释率和累积总共方差解释率。
- 分析loading载荷系数值
 - 通过因子载荷系数值（经验阈值0.3），分析出每个因子与题项的对应关系情况；结合因子与题项对应关系，对各个因子进行命名。



Python因子分析：数据

- 使用了bfi2010的数据集，这个数据集收集了2800个人关于人格的25个问题，如：
 - Am indifferent to the feelings of others. (对别人的感受漠不关心)
 - Inquire about others' well-being. (询问他人是否幸福)
 - Know how to comfort others. (知道如何让安慰别人)
 - Love children. (喜欢孩子)
 - Make people feel at ease. (让人感到轻松)
 - Am exacting in my work. (对工作热情)
- 这些特征是与5个隐藏特征有关，这5个特征就是5大人格特征
- 关于特征之间的对应关系如下所示(之后的因子分析也是可以证明这个的)：
 - agree=c("A1","A2","A3","A4","A5") => 认同性
 - conscientious=c("C1","C2","C3","-C4","-C5") => 勤奋的, 责任感
 - extraversion=c("-E1","-E2","E3","E4","E5") => 外向的
 - neuroticism=c("N1","N2","N3","N4","N5") => 神经质, 不稳定性
 - openness = c("O1","-O2","O3","O4","-O5") => 开放性
- 其实我们是不知道这种对应关系的，这5大人格相当于5个隐藏变量，这些隐藏变量会导致我们观测的变化，所以我们相当于想要通过因子分析来找到这25个变量后面的隐藏变量。



Python因子分析

- 参见Jupyter Notebook文档

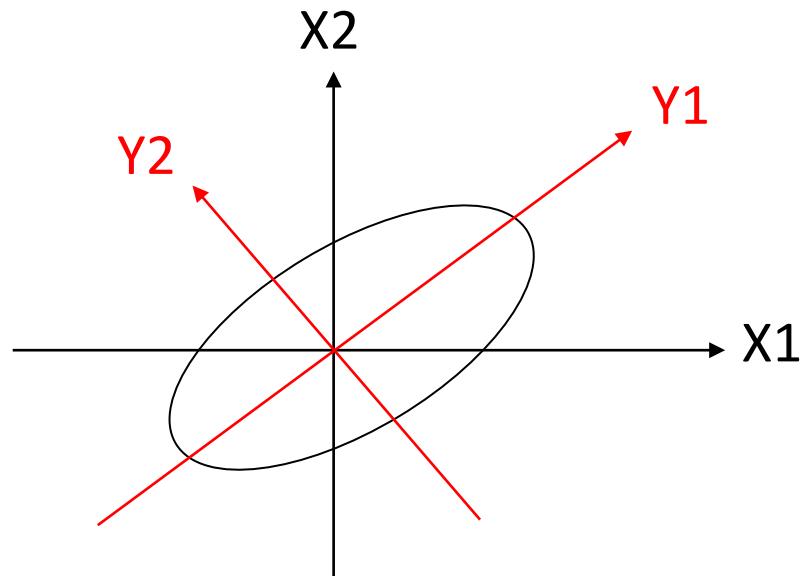




主成分分析法 (Principal Component Analysis, PCA)

- 假设记录包括n个字段，挑选最能表示数据变异的k个维度的正交向量，产生维度的缩减
 - 仅是计算维度的减少，数据输入的维度则未改变

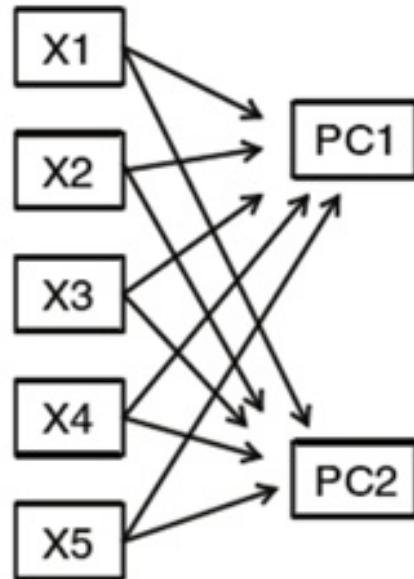
若原本数据集有 k 个变量(即有 k 个维度)，利用线性变换的方式找到 c 个新的变量($c \leq k$)来表示原本的资料



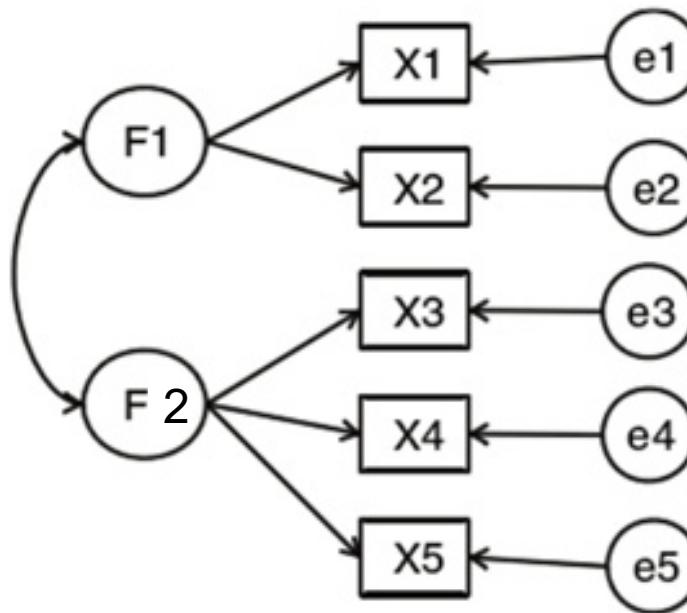


主成分分析和因子分析的区别

- 主成分分析，是分析维度属性的主要成分表示；
因子分析，是分析属性们的公共部分的表示。



(a) Principal Components Model



知乎 @syf写字的地方
(b) Factor Analysis Model



主成分分析法（PCA）

- 将m个变量进行线性组合以得到n个新的变量，这些新变量称为主成分

$$P_1 = a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + \cdots + a_{1m}X_m$$

$$P_2 = a_{21}X_1 + a_{22}X_2 + a_{23}X_3 + \cdots + a_{2m}X_m$$

⋮

$$P_m = a_{m1}X_1 + a_{m2}X_2 + a_{m3}X_3 + \cdots + a_{mm}X_m$$

- 假设有m个原始变量，则最多可求得m个主成分
- 如何选择n个主成分，需视对m个变量的解释能力





PCA计算范例 (1/2)

学生	1	2	3	4	5	6	7	8	9	10
英文修课成绩 X_1	90	88	60	73	82	69	77	88	59	72
英文检定成绩 X_2	280	261	183	277	230	234	263	278	197	230

学生在校英语成绩与检定成绩的协方差矩阵

$$\mathbf{S} = \begin{bmatrix} 126.62 & 316.84 \\ 316.84 & 1178.68 \end{bmatrix}$$

假设特征值为 $\lambda_1 \quad \lambda_2$

则满足 $\det(\mathbf{S} - \lambda \mathbf{I}) = \begin{vmatrix} 126.62 - \lambda & 316.84 \\ 316.84 & 1178.68 - \lambda \end{vmatrix} = 0$

计算得 $\lambda_1 = 1266.73 \quad \lambda_2 = 38.57$

假设对应的特征向量为 $a_1 = \begin{pmatrix} a_{11} \\ a_{12} \end{pmatrix} \quad a_2 = \begin{pmatrix} a_{21} \\ a_{22} \end{pmatrix}$

也满足 $\mathbf{S} \cdot \begin{pmatrix} a_{11} \\ a_{12} \end{pmatrix} = \lambda_1 \cdot \begin{pmatrix} a_{11} \\ a_{12} \end{pmatrix} \Rightarrow \begin{pmatrix} a_{11} \\ a_{12} \end{pmatrix} = \begin{pmatrix} 0.268 \\ 0.963 \end{pmatrix} \Rightarrow P_1 = 0.268X_1 + 0.963X_2$

$\mathbf{S} \cdot \begin{pmatrix} a_{21} \\ a_{22} \end{pmatrix} = \lambda_2 \cdot \begin{pmatrix} a_{21} \\ a_{22} \end{pmatrix} \Rightarrow \begin{pmatrix} a_{21} \\ a_{22} \end{pmatrix} = \begin{pmatrix} 0.963 \\ -0.268 \end{pmatrix} \Rightarrow P_2 = 0.963X_1 - 0.286X_2$





PCA计算范例 (2/2)

- 利用协方差矩阵或者相关系数矩阵进行分析
 - 当单位不同或将个别变量的权重视为相同时，则使用相关矩阵分析较适当
- 假设修课成绩最高为100分，检定成绩最高为300分
 - 两变数的总变异中 X_2 会占大多数而使得萃取出的主成分偏向 X_2
 - 要避免因单位不同或变异数差异过大造成变数间权重不一

相关系数矩阵为 $\mathbf{R} = \begin{pmatrix} 1.0 & 0.82 \\ 0.82 & 1.0 \end{pmatrix}$

则满足 $\det(\mathbf{R} - \lambda \mathbf{I}) = \begin{vmatrix} 1.0 - \lambda & 0.82 \\ 0.82 & 1.0 - \lambda \end{vmatrix} = 0$

计算得 $\lambda_1 = 1.82$ $\lambda_2 = 0.18$

也求得 $P_1 = 0.707X_1 + 0.707X_2$ $P_2 = 0.707X_1 - 0.707X_2$



PCA的结果与解释 (1/5)

1. 主成分分析是否合适

- 若为使得新变量彼此完全不相关，以作进一步分析，则需看主成分是否可解释或具有特殊意义
- 如果难以解释或没有意义，则不建议利用主成分产生新变量

2. 主成分能解释的方差百分比

- 主成分只取特征值最大的前n个来代替原来的m个变量， $V(P_j) = \lambda_j$ 而所有主成分的总方差等于所有原始变量的总方差

$$\sum_{j=1}^m V(P_j) = \sum_{i=1}^m V(X_i)$$

- 对于第j个主成分解释方差的比例，即为第j个主成分的特征值占所有特征值总和的比例

$$\frac{V(P_j)}{\sum_{j=1}^m V(P_j)} = \frac{\lambda_j}{\sum_{i=1}^m V(X_i)} = \frac{\lambda_j}{\sum_{j=1}^m \lambda_j}$$





PCA的结果与解释 (2/5)

- 标准化后的主成分特征值总和等于原有变数个数和

$$\sum_{j=1}^m \lambda_j = m$$

- 若选取n个主成分则会解释原有m个变量的变异比例

$$\sum_{j=1}^n \lambda_j \Bigg/ \sum_{j=1}^m \lambda_j$$

- 以修课成绩 (X_1) 与检定成绩 (X_2) 为例

- 第一主成分解释的变异百分比为: $1266.73/(1266.73 + 38.57) = 97.05\%$
- 第二主成分解释的变异百分比为: $38.57/(1266.73 + 38.57) = 2.95\%$

- 若相关系数矩阵，则

- 第一主成分解释的变异百分比为: $1.82/(1.82 + 0.18) = 91.00\%$
- 第二主成分解释的变异百分比为: $0.18/(1.82 + 0.18) = 9.00\%$

- 为简化变量，在可接受的损失下，可以利用少数主成分来代表原有数据的变异



PCA的结果与解释 (3/5)

3. 主成分个数的选取

- 取决于可容忍的信息损失比例

- 若选取越多，涵盖原有信息变异的程度就越高
- 若选取较少，可能导致主成分的代表性不足

- 常用的主成分选取方法

- 对于标准化的数据，仅选取特征值大于1的主成分 $\lambda \geq 1$
- 利用陡坡图(scree plot)将每个主成分排序，并对应特征值
- 变异数的解释比例比需超过某一水平（通常为0.7, 0.85, 0.95 以上）

- 唯有变量高度相关时，才能以少数主成分来代替原有变数
- 变量间相关性越强，需要的主成分个数越少



PCA的结果与解释 (4/5)

4. 主成分的解释

- 主成分负荷(loadering)可判断和解释主成分的构成

$$l_{ij} = \frac{a_{ij}\sqrt{\lambda_i}}{s_j}$$

s_j 为第j个原始变量的标准差

以负荷大于0.5作为该变量对主成分为有影响的准则

- 以修课成绩(X_1)与检定成绩 (X_2)为例

若输入为协方差矩阵，则其负荷计算分别为：

$$l_{11} = \frac{a_{11}\sqrt{\lambda_1}}{s_1} = \frac{0.268\sqrt{1266.73}}{11.25} = 0.848$$

$$l_{12} = \frac{a_{12}\sqrt{\lambda_1}}{s_2} = \frac{0.936\sqrt{1266.73}}{34.33} = 0.998$$

$$l_{21} = 0.532 \quad l_{22} = -0.048$$

若输入为相关系数矩阵，则其负荷计算分别为：

$$l_{11} = a_{11}\sqrt{\lambda_1} = 0.707 \times \sqrt{1.82} = 0.954$$

$$l_{12} = a_{12}\sqrt{\lambda_1} = 0.707 \times \sqrt{1.82} = 0.954$$

$$l_{21} = a_{21}\sqrt{\lambda_2} = 0.707 \times \sqrt{0.18} = 0.300$$

$$l_{22} = a_{22}\sqrt{\lambda_2} = -0.707 \times \sqrt{0.18} = -0.300$$





PCA的结果与解释 (5/5)

- 主成分的应用

- 当分析需要对一组变量订定出一个总指标（或指数）并给定权重时，主成分能由资料推估找出最有解释能力的变量组合和对应的解释比重，也可以作为发展综合指数的参考

- 降维
- 去除相关性



使用Python进行主成分分析

- 参见Jupyter Notebook文档





布置期中练习（15分）

- 截止日期：5月10日15:09
- 个人完成。





谢谢！

