小象学院
ChinaHadoop.cn

第五课（第13-15课时）

探索变量之间的关系

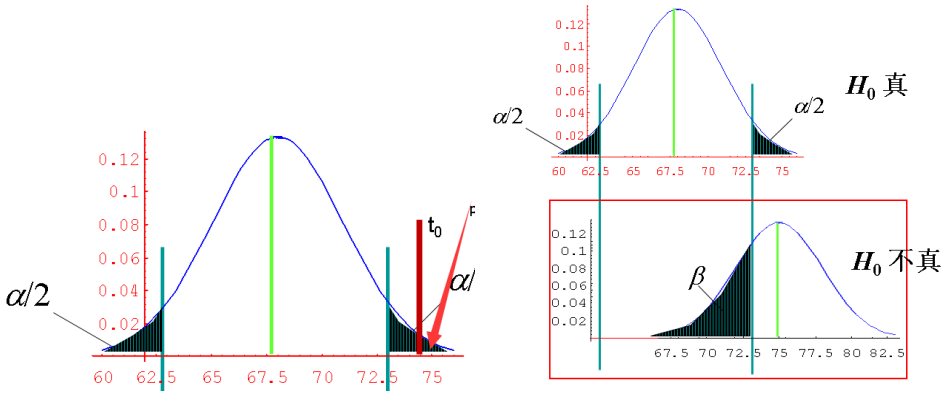
- 图形分析
- 相关分析
- 方差分析

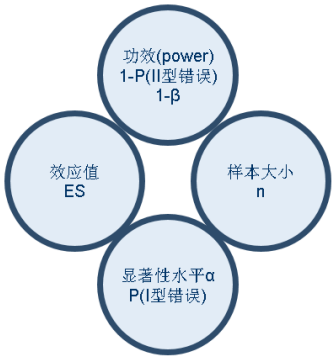
前情回顾

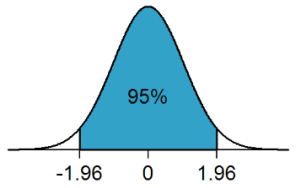
小象学院 ChinaHadoop.cn

- **如何检验一个变量的一组取值是否符合某种分布**
  - 图形分析
  - 使用样本数字特征
  - 使用
- **什么是假设检验？**
  - 统计分析的经典框架
  - p-值
  - 从样本分布构建检验
  - K-S检验
  - 以及其他分布检验

		判断	
		拒绝H0	接受H0
真实	H0为真	I型错误	正确
	H0为假	正确	II型错误







任务描述

- 理解变量之间的关系
- 了解根据变量类型，选择适合的分析方式
- 掌握使用图形分析，相关分析，方差分析

- 数据：链接: http://pan.baidu.com/s/1bpKAd8V 密码: dw8g
  - tips.csv

载入数据：tips.csv

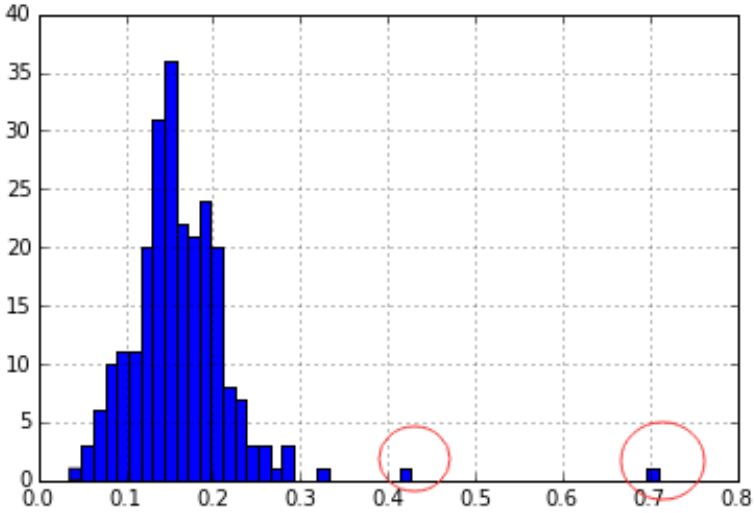
小象学院 ChinaHadoop.cn

- **载入常用库**
  - import pandas as pd
  - import numpy as np
  - import matplotlib.pyplot as plt
- **载入模块**
  - from pandas import Series, DataFrame
  - from scipy import stats
- **读入数据（假设文件在工作目录路径下）**
  - tips=pd.read_csv('tips.csv')

从单变量到多变量

小象学院 ChinaHadoop.cn

- **对单变量的分析：**
  - 数字特征
  - 分布
    - “模式”？
    - “异常”？



- **对两个或更多变量的分析：**
  - 找到变量之间的关系……
  - case “没关系”：
    - Nothing more to do
  - case “有关系”：
    - 多大关系？
    - 什么样的关系？

设随机事件A, B满足 $P(AB)=P(A)P(B)$
则称事件A与B相互独立，简称A与B独立.

设A, B为两个事件，$P(A)P(B)>0$，则
A与B独立的充分必要条件是
$P(A|B)=P(A)$或$P(B|A)=P(B)$.

事件A与B相互独立,是指其中任一事件发生的概率都不受另外一事件发生的影响.

度量变量之间的关系

小象学院 ChinaHadoop.cn

- **使用可视化**
  - 探索变量之间的关系
  - 点图
  - 箱线图
- **独立性检验**
  - 变量之间是否相互独立
- **相关性检验**
  - 独立性检验被拒绝时，考察相关性
  - 相关系数

可视化分析-使用点图

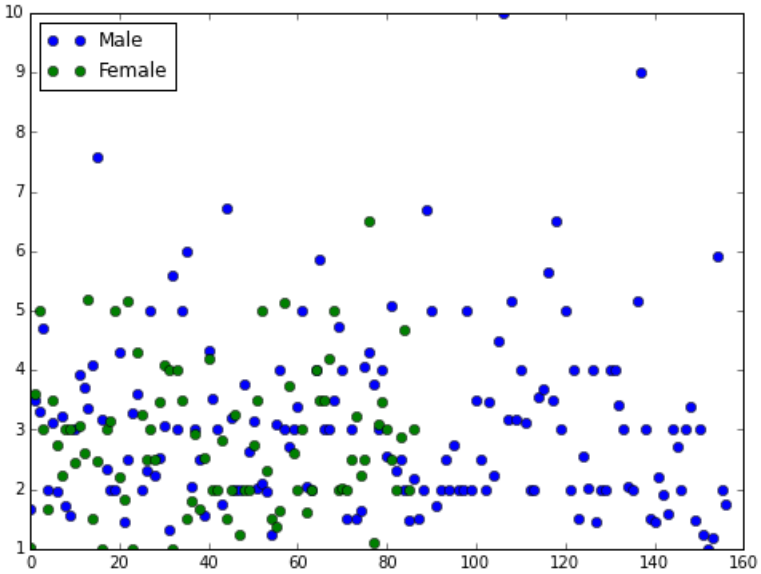
小象学院 ChinaHadoop.cn

- fig=plt.figure(figsize=(8,6))
- ax=fig.add_subplot(1,1,1)
- from numpy import random
- ax.plot(random.rand(50).cumsum(),'.')
- fig

可视化分析-使用点图

小象学院 ChinaHadoop.cn

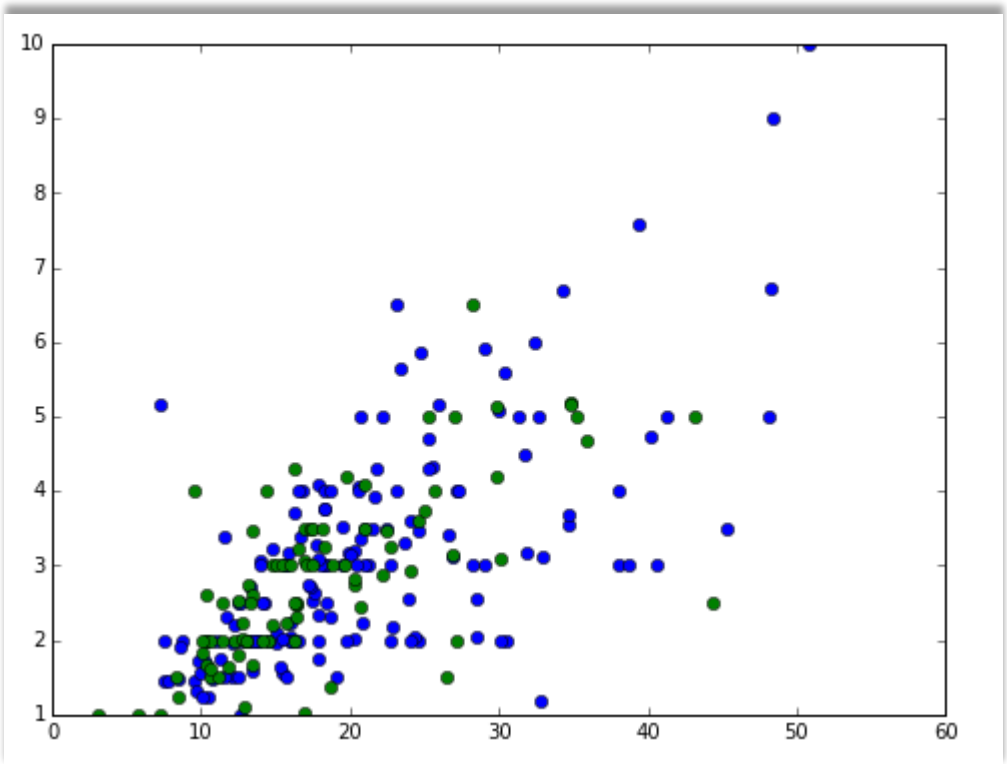
- fig,ax=plt.subplots(1,1,figsize=(6,6))
- ax.clear()
- ax.plot(tips[tips['sex']=='Male']['tip'],'o',label='Male')
- ax.plot(tips[tips['sex']=='Female']['tip'],'o',label='Female')
- ax.legend(loc='best')
- fig



可视化分析-使用点图

小象学院 ChinaHadoop.cn

- **使用散点图，分组**
  - ax.plot(tips[tips['sex']=='Male']['total_bill'],tips[tips['sex']=='Male']['tip'],'o',label='Male')
  - ax.plot(tips[tips['sex']=='Female']['total_bill'],tips[tips['sex']=='Female']['tip'],'o',label='Female')
  - fig



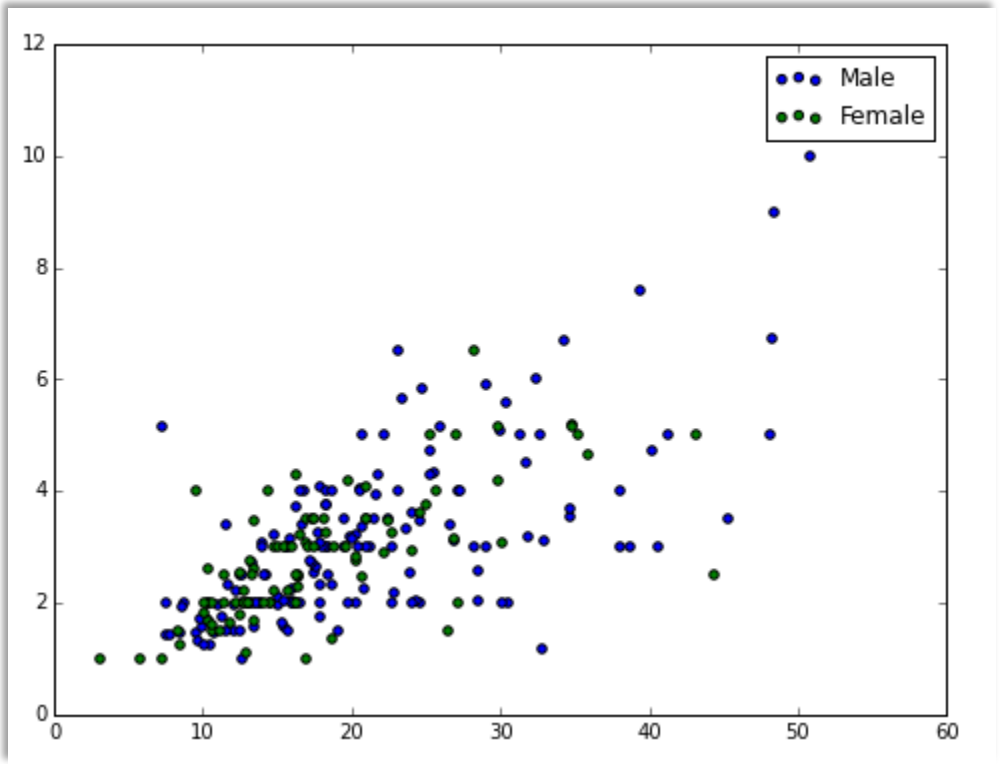
可视化分析-使用点图

- **可视化的优势**
  - 更容易发现“模式”

使用散点图

小象学院 ChinaHadoop.cn

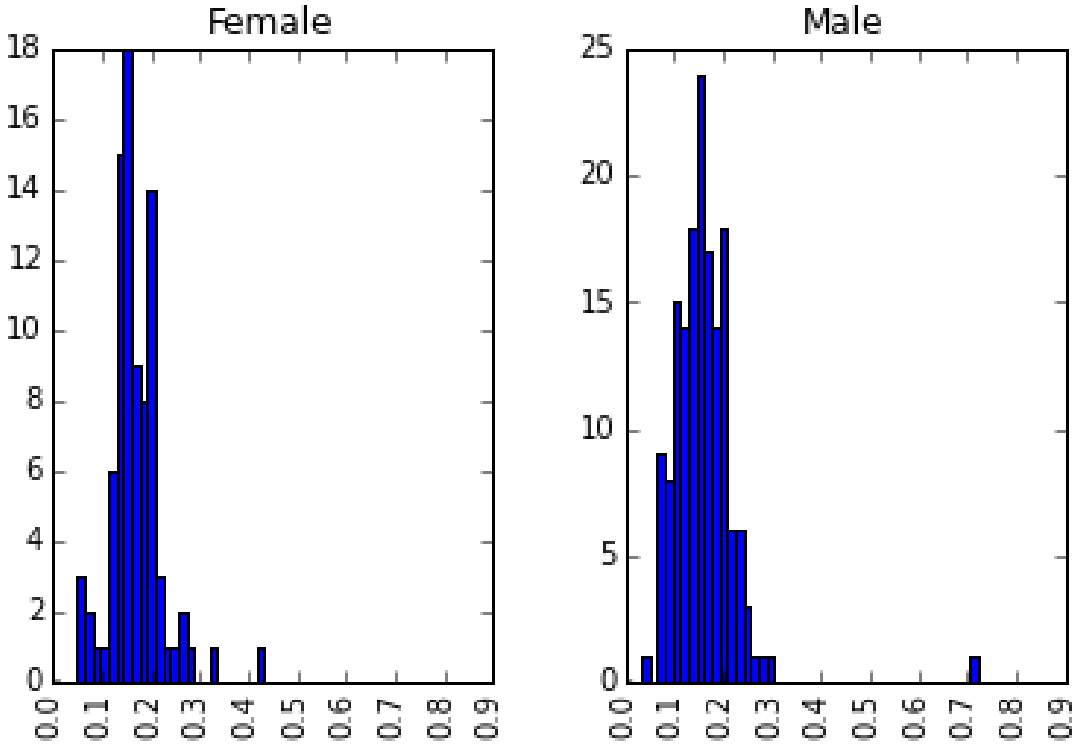
- **分组**
  - ax.scatter(tips[tips['sex']=='Male']['total_bill'],tips[tips['sex']=='Male']['tip'],label='Male')
  - ax.scatter(tips[tips['sex']=='Female']['total_bill'],tips[tips['sex']=='Female']['tip'],c='g',label='Female')
  - ax.legend(loc='best')
  - fig



可视化分析-使用点图-继续探索

小象学院 ChinaHadoop.cn

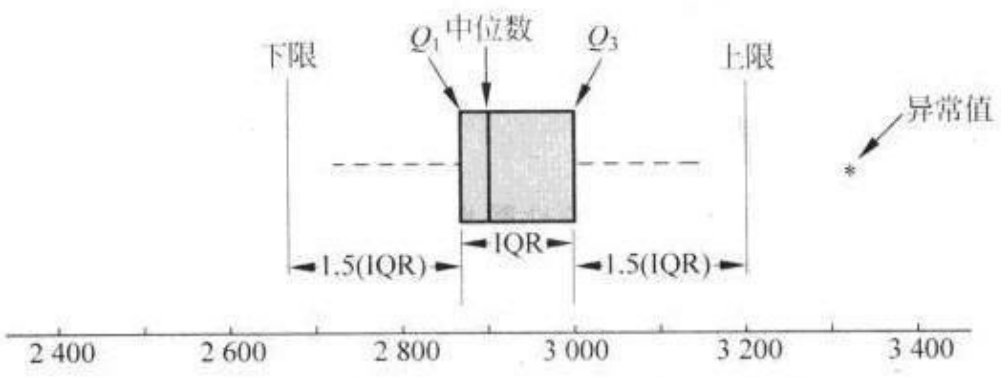
- **直方图**
  - tips['tip_pct']=tips['tip']/tips['total_bill']
  - tips['tip_pct'].hist(by=tips['sex'],bins=50,range=[0,0.8])

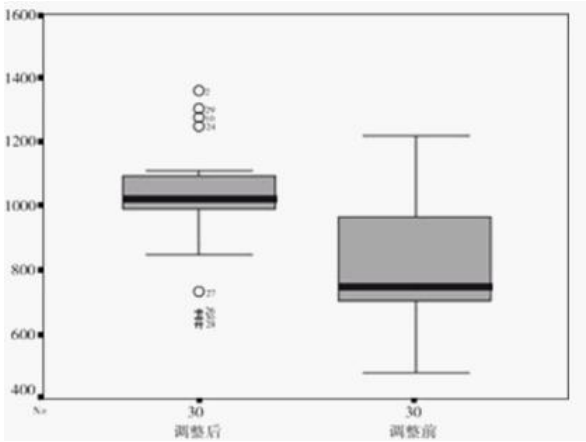


可视化分析-使用箱线图

小象学院 ChinaHadoop.cn

- **箱线图（盒须图）**
  - 中位数
  - 四分位数全距$IQR = Q_3 - Q_1$
  - 箱线图的上、下限制线分别在比Q_1低1.5(IQR)和比Q_3高1.5(IQR)的位置上
  - 异常值

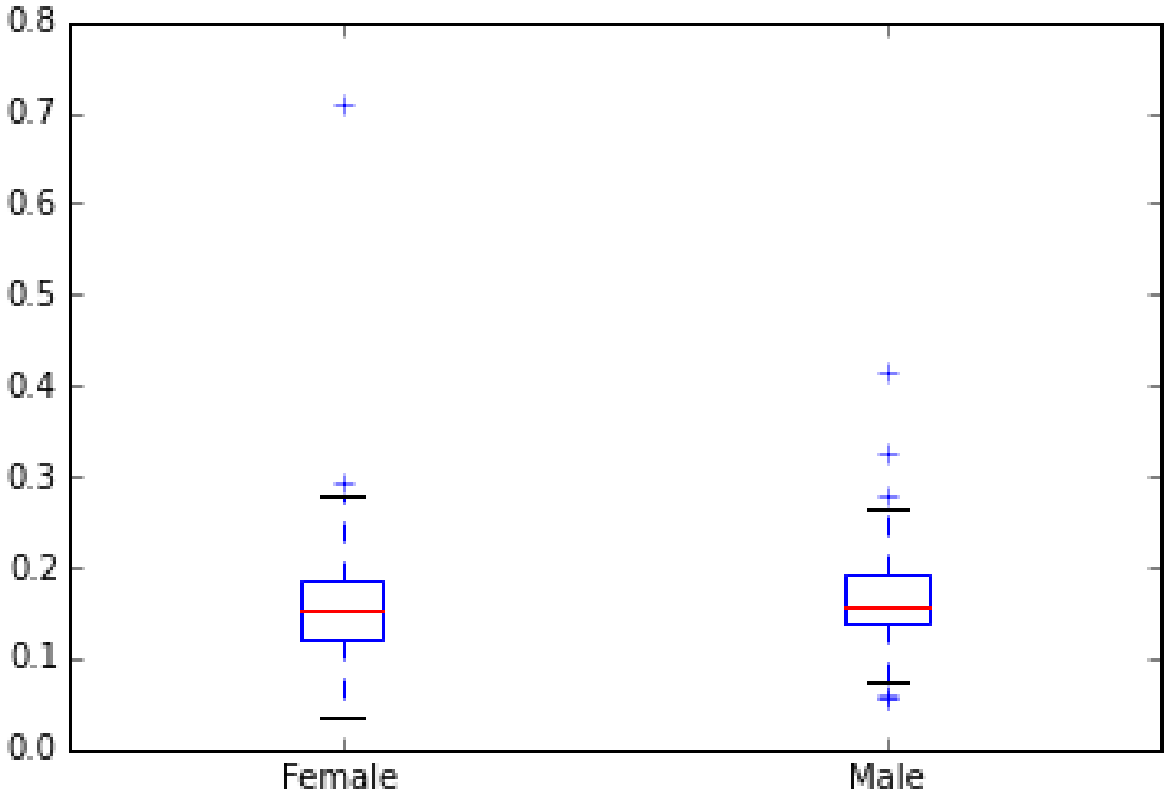




可视化分析-使用箱线图

小象学院 ChinaHadoop.cn

- **boxplot()**
  - rvs1=tips[tips['sex']=='Male']['tip_pct']
  - rvs2=tips[tips['sex']=='Female']['tip_pct']
  - plt.boxplot([rvs1,rvs2],labels=['Female','Male'])



进一步探索变量之间的关系

- 可视化给出可能的方向
- 需建立更严格的分析方式：
  - 假设检验
- 变量的类型
  - 类别型
  - 数值型

属性类型	表述和允许的变换	例子	操作
Nominal 标称 （分类、定性）	与其他对象区别的名称 (=, ≠) 一对一变换	邮编、ID、姓名、性别	众数, 熵, 列联相关 χ^2 检验
Ordinal 序数 （分类、定性）	确定对象信息的序. (<, >) 保序变换	矿石硬度、成绩、街道号码	中值，百分位，秩相关
Interval 区间 （数值、定量）	区间属性，值之间的和差是有意义的 (+, -) 线性变换	日期，温度	均值、标准差、Pearson相关
Ratio 比例 （数值、定量）	比率变量，积和比有意义 (*, /) 线性变换（乘积）	绝对零度、货币量、计数、年龄	几何平均，调和平均，百分比变差

度量变量之间的关系

小象学院 ChinaHadoop.cn

- **然而，实际中问题有不同的提出方式：**
  - 星座会决定性格吗？
    - 变量“星座”vs 变量“性格”
  - 游客对不同品牌的酒店评价差别大吗？
    - 变量“评价”vs变量“酒店品牌”
  - 男女性别中给小费的费率是否不同？
    - 费率~性别
  - 小费和总花费之间是否相关？
    - 小费~总花费
  - 星期几是否会影响每组就餐的人数？
    - 星期几~每组就餐人数

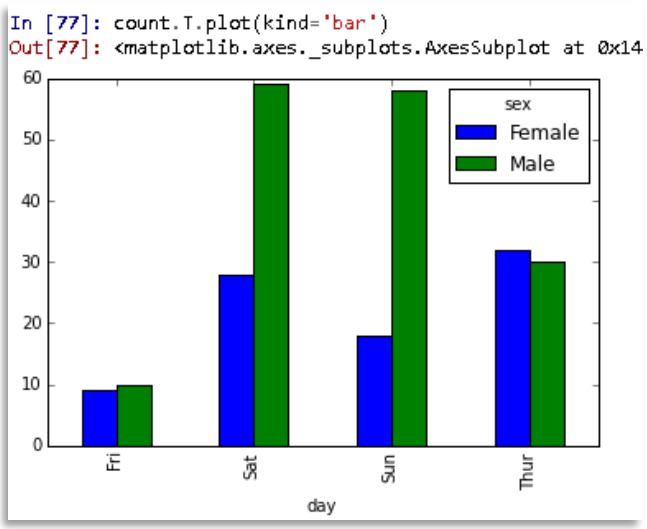
独立性检验

- **类别型~类别型**
- **星期几跟是否吸烟有关系吗？**
  - 当然，提问题的时候先画图
    - count=pd.crosstab(tips.sex, tips.day)
    - count.T.plot(kind='bar')
  - 然后用定量方法确定答案
    - chi2, p, dof, ex = stats.chi2_contingency(count, correction=False)
    - p

```
In [27]: count
Out[27]:
day      Fri  Sat  Sun  Thur
sex
Female     9   28   18    32
Male      10   59   58    30
```

Contingency Table:列联表

注意：每个cell中的数至少为5



独立性检验：为什么是卡方检验？

小象学院 ChinaHadoop.cn

- **复习知识点**
  - 联合分布
  - 边缘分布

X \ Y	y_1	y_2	$\cdots$	y_j	$\cdots$	$P\{X=x_i\}$
x_1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$	$\sum_j p_{1j}$
x_2	p_{21}	p_{22}	$\cdots$	p_{2j}	$\cdots$	$\sum_j p_{2j}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
x_i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$	$\sum_j p_{ij}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
$P\{Y=y_j\}$	$\sum_i p_{i1}$	$\sum_i p_{i2}$	$\cdots$	$\sum_i p_{ij}$	$\cdots$	

两事件A,B独立的定义是：
若P(AB)=P(A)P(B)
则称事件A,B独立 .

独立性检验

小象学院 ChinaHadoop.cn

- **类别~类别**
- **费舍尔精确检验（如果样本较小）**
  - count=pd.crosstab(tips.sex, tips.smoker)
  - oddsratio, pvalue = stats.fisher_exact(count)
  - (1.012183692596063 7, 1.0)

  - count.iat[0,0]=2
  - stats.fisher_exact(count)
  - (0.037488284910965321, 3.9900059898475714e-10)

```
In [35]: count
Out[35]:
smoker  No  Yes
sex
Female   2   33
Male    97   60
```

独立性检验

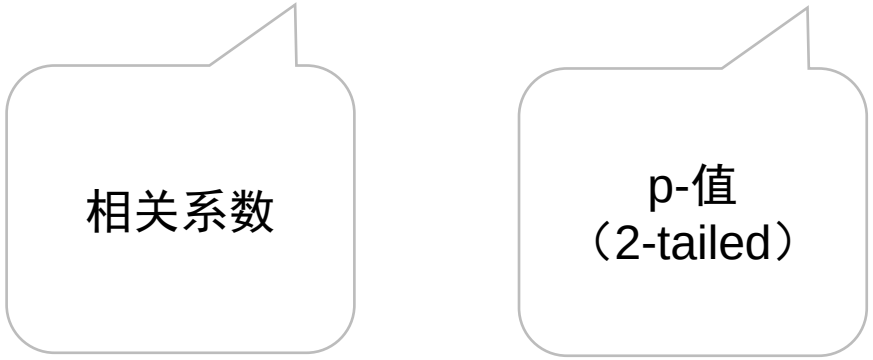
小象学院 ChinaHadoop.cn

- **数值型~数值型**
- **相关度**
  - Pearson：积差相关系数，反应两变量之间线性相关性
  - Spearman：等级相关系数（Ranked data）
  - Kendall's Tau：非参数等级相关系数

独立性检验

小象学院 ChinaHadoop.cn

- 数值型~数值型
- Pearson相关系数
- stats.pearsonr
  - stats.pearsonr(tips.total_bill,tips.tip)
  - Out[25]: (0.6757341092113643, 6.692470646864040 7e-34)

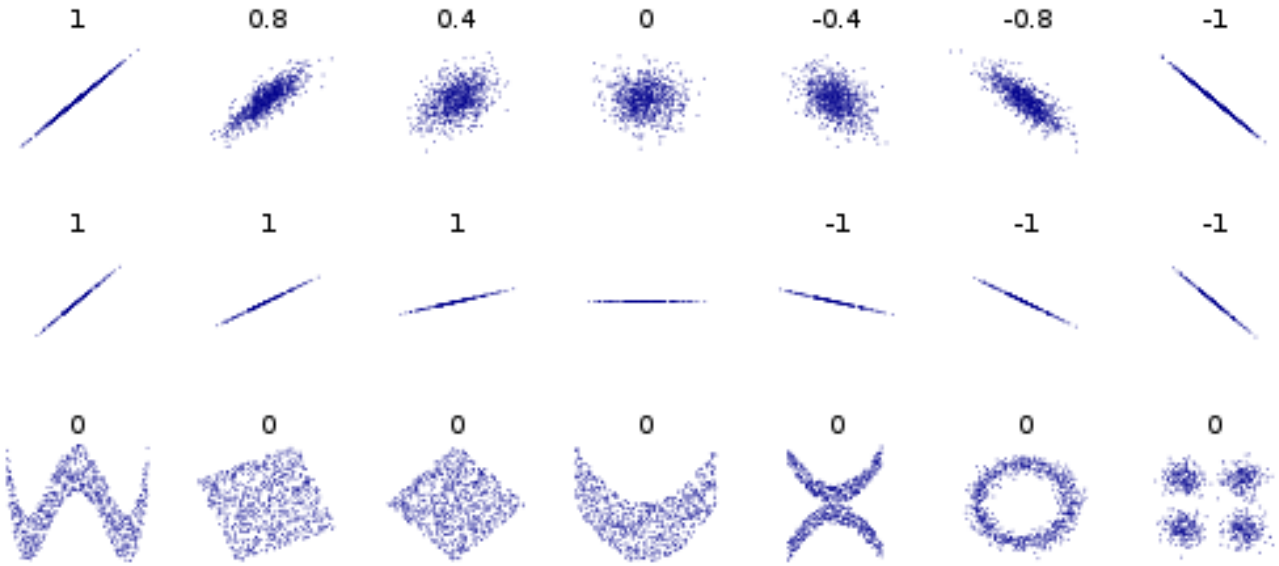


独立性检验

小象学院 ChinaHadoop.cn

- **Pearson相关系数**

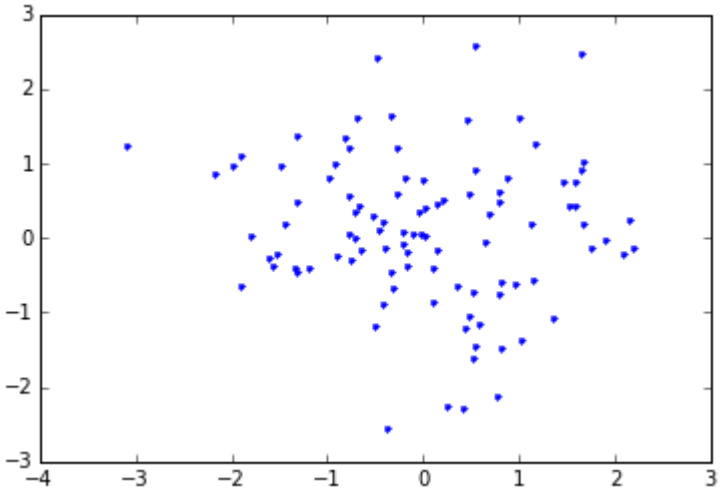
$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$



独立性检验

小象学院 ChinaHadoop.cn

- **数值型~数值型**
- **Spearman相关系数：**
  - 与pearson相关系数不同，不假设变量为正态分布
  - -1~+1 衡量变量之间的单调性
  - np.random.seed(1234321)
  - x = np.random.randn(100)
  - y = np.random.randn(100)
  - rho, pval = stats.spearmanr(x, y)



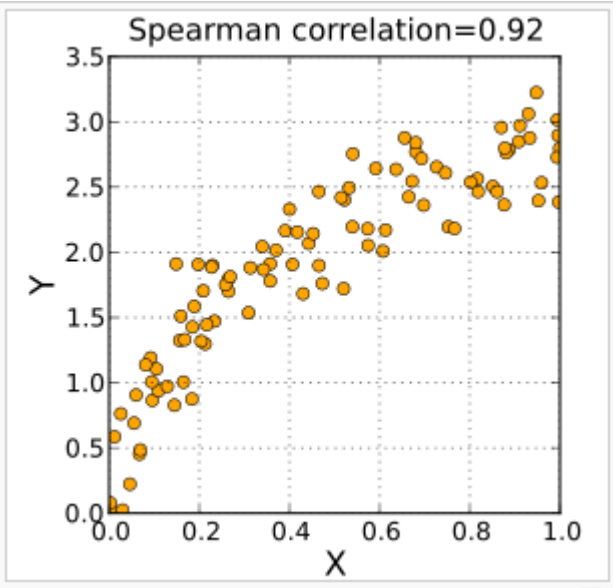

```
In [70]: stats.pearsonr(x,y)
Out[70]: (-0.095440544065648725, 0.34488119203969791)

In [71]: stats.spearmanr(x,y)
Out[71]: (-0.095349534953495352, 0.34534352767739207)
```

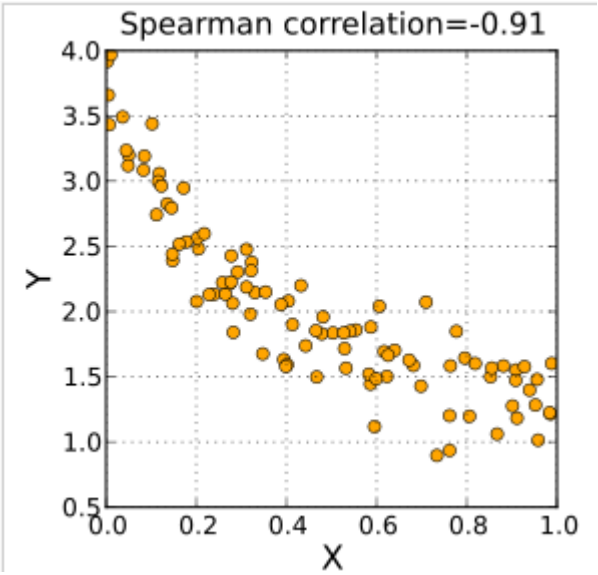
独立性检验

小象学院 ChinaHadoop.cn

- **Spearman相关系数**

$$r_s = \rho_{\mathrm{rg}_X, \mathrm{rg}_Y} = \frac{\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X} \sigma_{\mathrm{rg}_Y}}$$





独立性检验

小象学院 ChinaHadoop.cn

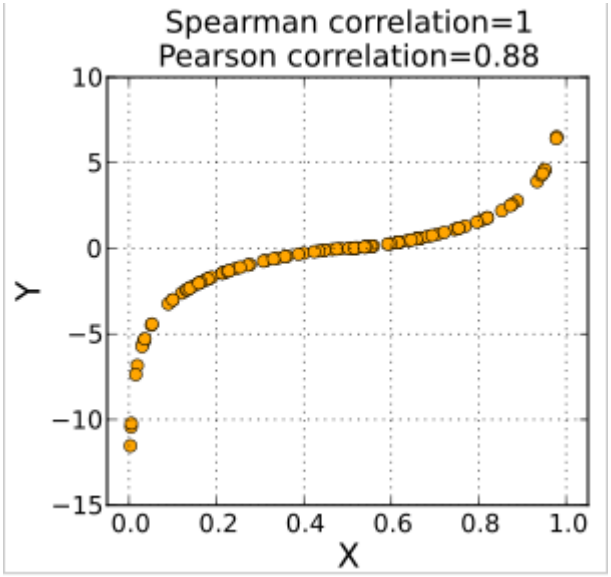
- **Spearman相关检验**
  - y2n = np.random.randn(100,2)
  - x2n = np.random.randn(100,2)
  - rho, pval = stats.spearmanr(x2n, y2n)
  - rho, pval = stats.spearmanr(x2n.T, y2n.T, axis=1)

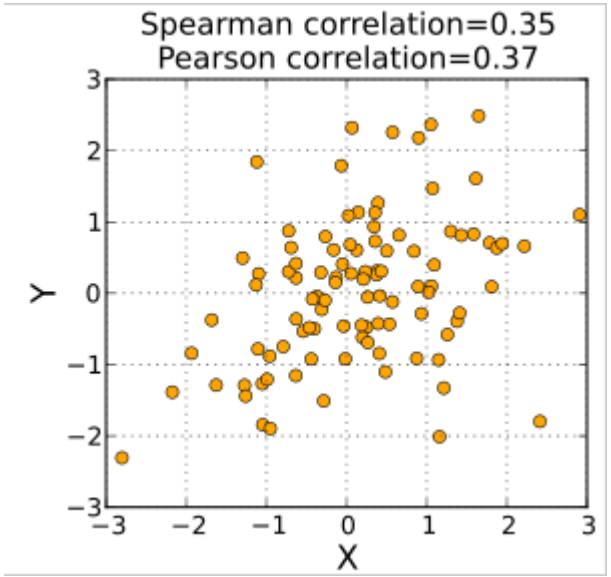
```
In [8]: rho
Out[8]:
array([[ 1.        , -0.02234623, -0.01081308,  0.0060366 ],
       [-0.02234623,  1.        ,  0.06233423,  0.05005701],
       [-0.01081308,  0.06233423,  1.        , -0.1240084 ],
       [ 0.0060366 ,  0.05005701, -0.1240084 ,  1.        ]])

In [9]: pval
Out[9]:
array([[ 0.        ,  0.82534128,  0.91496792,  0.95246815],
       [ 0.82534128,  0.        ,  0.53782525,  0.62089176],
       [ 0.91496792,  0.53782525,  0.        ,  0.21898027],
       [ 0.95246815,  0.62089176,  0.21898027,  0.        ]])
```

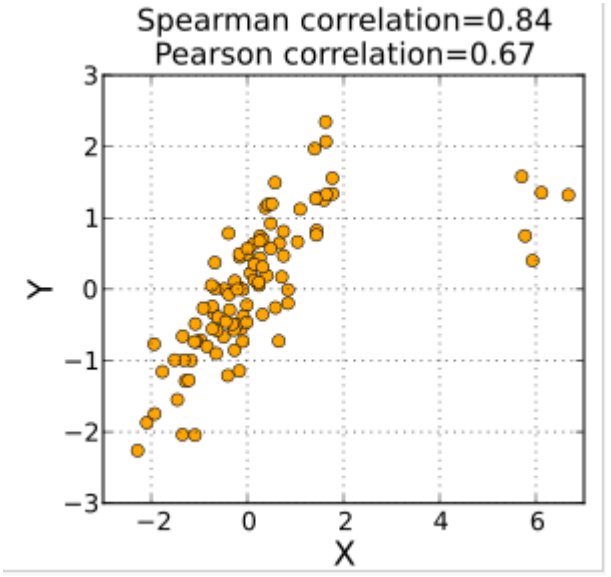
独立性检验（相关性检验）

小象学院 ChinaHadoop.cn

- **Pearson v.s. Spearman**







独立性检验

小象学院 ChinaHadoop.cn

- **数值型~数值型**
- **Kendall's Tau相关系数**
- **tau = (P - Q) / sqrt((P + Q + T) * (P + Q + U))**
  - p：同步数据对数，Q：异步，T：tie in x, U: tie in y
  - x=[1, 2, 3, 4, 5, 5, 4, 6,-1]
  - y=[3, 4, 3, 4, 5, 5, 3, 7, 7]
  - tau, p_value = stats.kendalltau(x,y)
  - (0.0468686868686869, 0.489613834011271)

```
In [12]: stats.kendalltau(x,y)
Out[12]: (0.0468686868686869, 0.489613834011271)

In [13]: stats.spearmanr(x,y)
Out[13]: (0.06313831383138314, 0.5325827936476879)

In [14]: stats.pearsonr(x,y)
Out[14]: (0.07021298609488854, 0.4875820383022492)
```

独立性（相关性）检验

小象学院 ChinaHadoop.cn

- **数值型~类别型**
- **男女性别中给小费的费率是否不同？**
  - 费率~性别
- **t检验**
  - rvs1=tips[tips['sex']=='Male']['tip']
  - rvs2=tips[tips['sex']=='Female']['tip']
  - stats.ttest_ind(rvs1, rvs2)

关联样本
stats.ttest_rel
（同组重复抽样）

独立性（相关性）检验

- **数值型~类别型**
- **男女性别中给小费的费率是否不同？**
  - 费率~性别
- **ANOVA方差分析**
- **ANCOVA协方差分析**
- **MANOVA多因素方差分析**

分组比较：方差分析（初步）ANOVA

- 因素效应的显著性分析： $y = \mu + \delta_i + \epsilon_{ij}$

$$H_0 : \delta_1 = \delta_2 = \ldots = \delta_a = 0 \leftrightarrow H_1 : \exists \delta_i \neq 0$$

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

$$SS_T = SS_E + SS_A$$

误差平方和 $$SS_E = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2$$

因素平方和 $$SS_A = \sum_{i=1}^{a} n_i (\bar{y}_{i\cdot} - \bar{y})^2$$

方差分析

小象学院 ChinaHadoop.cn

➢ 单因素方差分析（one-way ANOVA）

- 假设检验统计量和分布

$$F = \frac{SS_A/(a-1)}{SS_E/(n-a)} = \frac{MS_A}{MS_E}$$

$$F \sim F(a-1, n-a)$$

- p-value与显著水平α比较，p<α，拒绝H0；否则，接受H0

  - stats.f_oneway(rvs1,rvs2)
  - (1.173749551574689, 0.2797103849605855 3)
  - 用 F 假设检验：结果显著-->组间不同

方差分析

- **请用方差分析星期几对小费的比例是否有影响？**

方差分析

小象学院 ChinaHadoop.cn

- **方差分析检验对数据的假设**
  - 1. 样本之间相互独立
  - 2. 样本均来自正态分布
  - 3. 方差齐次性：各组方差相等

- **如何检验以上假设？**
  - 方差齐次性：stats.fligner(rvs1,rvs2)

```
In [21]: stats.fligner(rvs1,rvs2)
Out[21]: (0.95213740994947371, 0.32917583412198081)
```

- **如果不满足以上任意一条**
  - 怎么办？

ANOVA的非参数版本

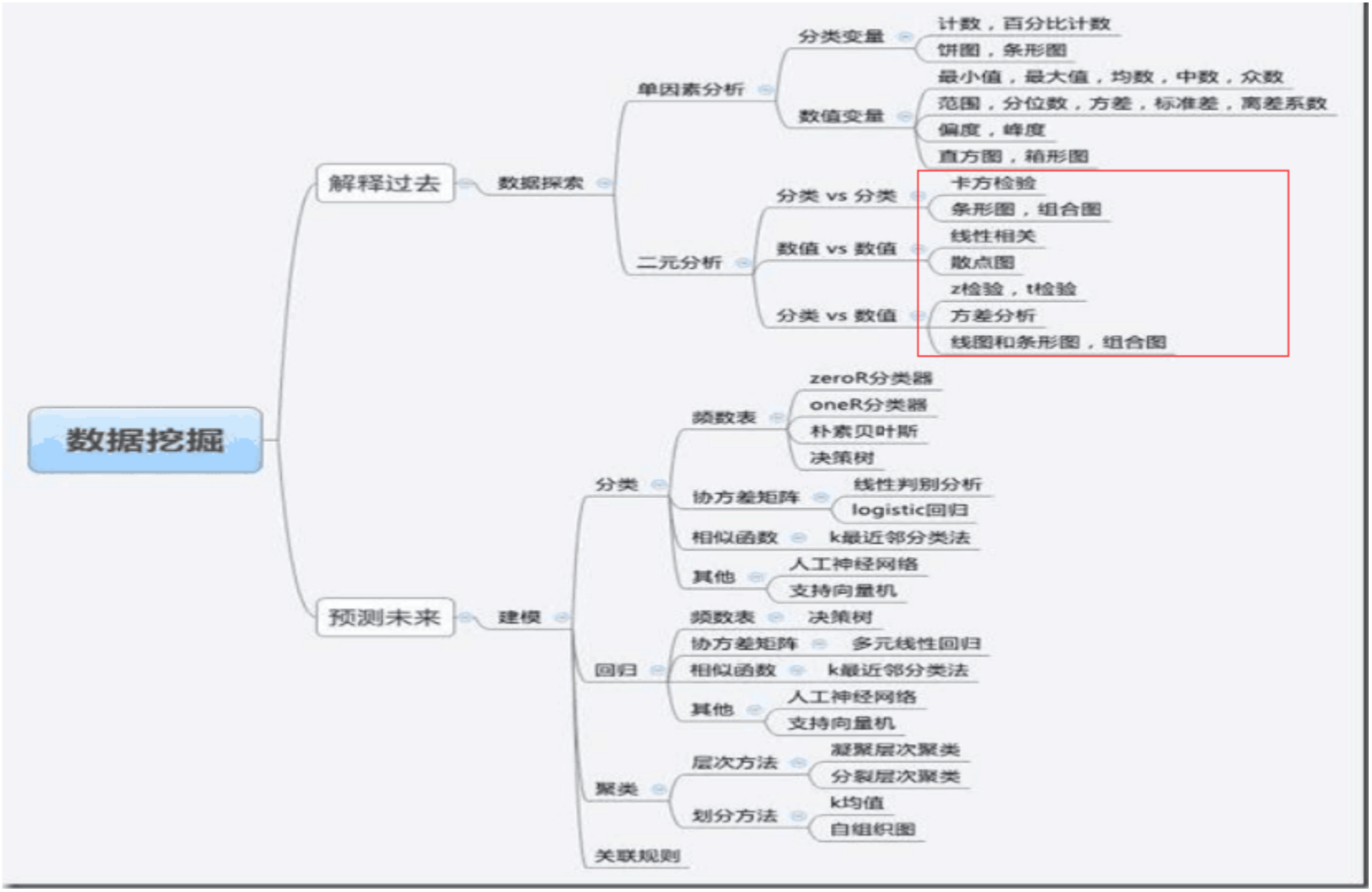
小象学院 ChinaHadoop.cn

- **Kruskal-Wallis H-test**
  - H0: 各组中值相等
  - 对数据亦有假设：Chi2分布，因此样本容量需不小于5
  - stats.kruskal(rvs1,rvs2)

```
In [20]: stats.kruskal(rvs1,rvs2)
Out[20]: (2.2351202029645436, 0.13490613264268328)
```

探索变量之间的关系

小象学院 ChinaHadoop.cn



由变量类型确定分析方式

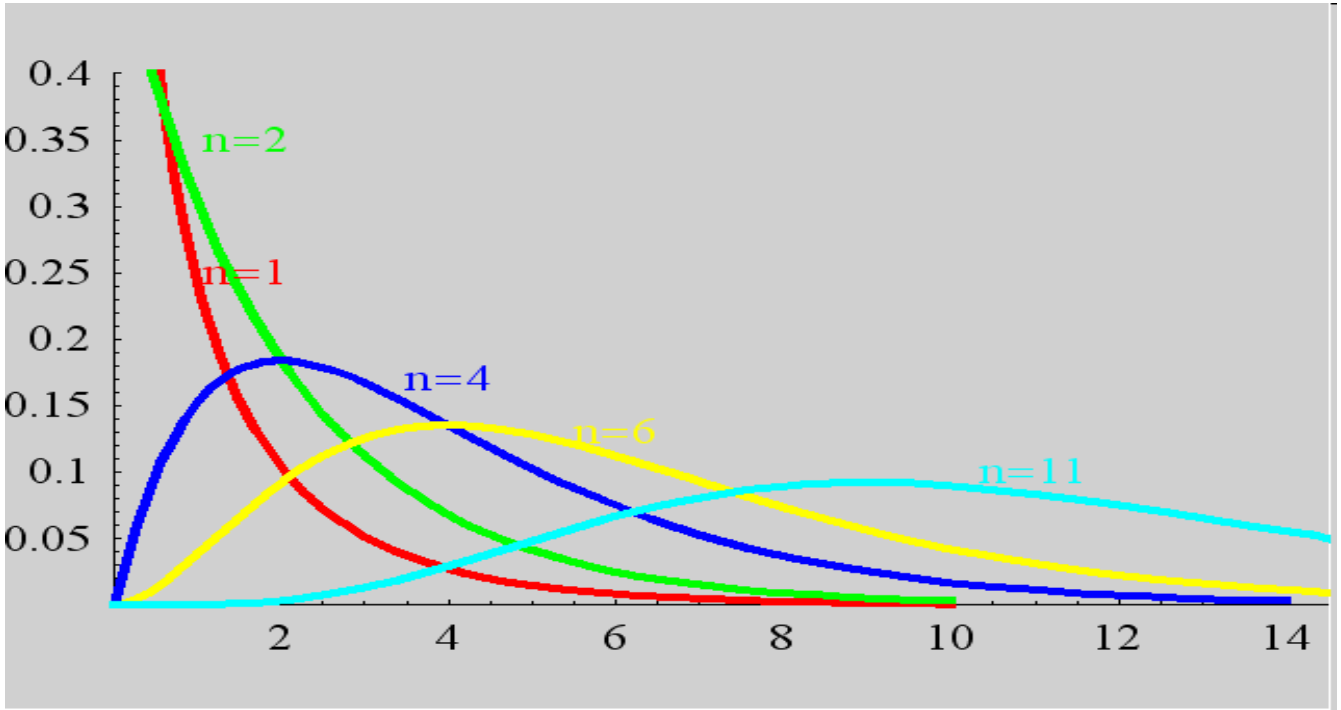
小象学院 ChinaHadoop.cn

- **然而，在这之前：变量需要构造**
  - 男女性别中给小费的费率是否不同？
    - 费率~性别
  - 小费和总花费之间是否相关？
    - 小费~总花费
  - 星期几是否会影响每组就餐的人数？
    - 星期几~每组就餐人数
  - 星座会决定性格吗？
    - 变量“星座”vs 变量“性格”
  - 游客对不同品牌的酒店评价差别大吗？
    - 变量“评价”vs变量“酒店品牌”

1.χ^2分布的定义

设 $X_1, X_2, \cdots, X_n$ 是来自总体 $X \sim N(0, 1)$ 的样本，

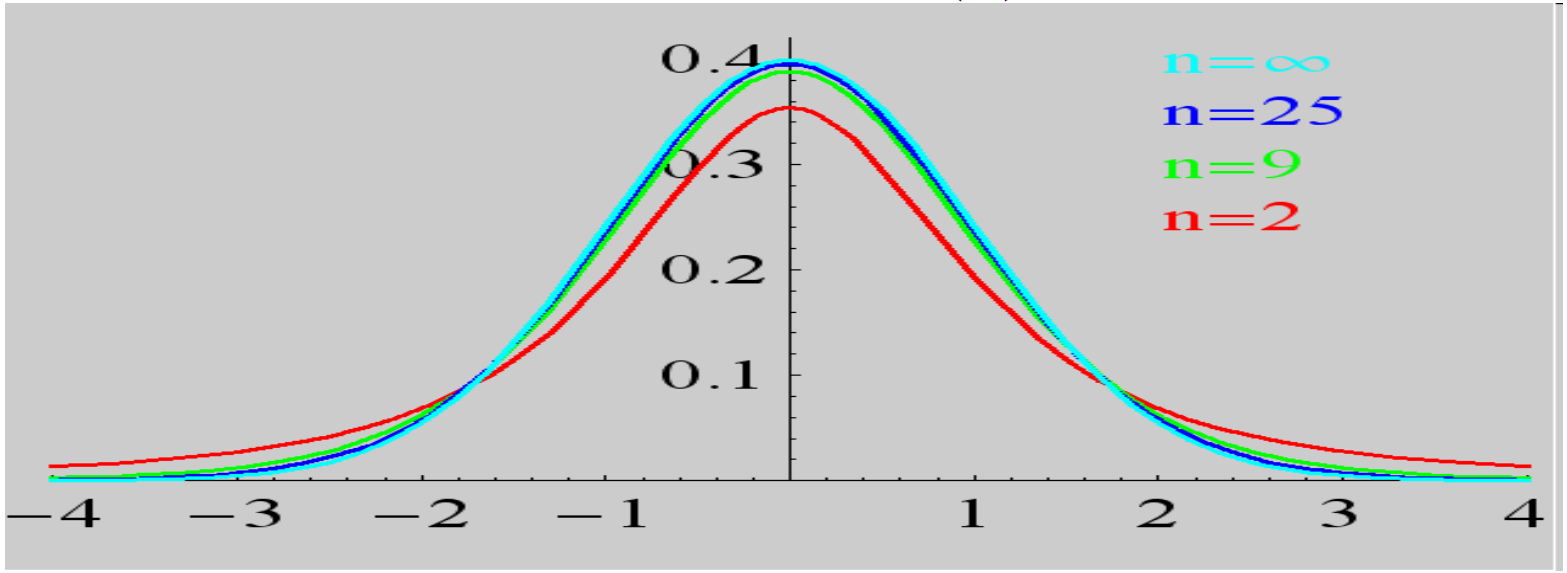
则 $X_1^2 + X_2^2 + \cdots + X_n^2 \sim \chi^2(n)$.



附：抽样分布

- **t分布**　设随机变量X服从标准正态分布$N(0,1)$，Y服从$\chi^2(n)$，且 X 与 Y 相互独立，

记 $T=\dfrac{X}{\sqrt{Y/n}}$，则 随机变量T服从自由度为n的t分布，记作$T\sim t(n)$.



附：抽样分布

➢ *t* 分布

设总体 $X \sim N(\mu,\sigma^2)$, $X_1, X_1, \cdots, X_n$（$n \geq 2$）是来自X的一个样本，$\bar{X}$ 与 S^2 分别为样本均值与样本方差，则随机变量

$$t = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

服从自由度为$n-1$的t分布.

附：抽样分布

小象学院 ChinaHadoop.cn

- ***F*分布**

设 $X \sim \chi^2(m), Y \sim \chi^2(n)$, 且 X 与 Y 相互独立，

记
$$Z = \frac{X/m}{Y/n} = \frac{nX}{mY}$$

则 Z 的密度函数为 $f(x;m,n)$, 因此 $Z \sim F(m,n)$.

由定理5.3.4不难看出，若 $X \sim F(m,n)$，则 $X^{-1} \sim F(n,m)$.

附：抽样分布

➢ *F*分布

$X_1, X_1, \cdots, X_{n_1}$和$Y_1, Y_2, \cdots, Y_{n_2}$是分别来自两个相互独立的正态总体$N(\mu_1, \sigma_1^2)$和$N(\mu_2, \sigma_2^2)$的两个随机样本（$n_1$，$n_2 \geq 2$），则

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F(n_1 - 1，\ n_2 - 1)$$

特别，当$\sigma_1^2 = \sigma_2^2$时，统计量“两个样本方差之比”服从F分布，即

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1，\ n_2 - 1)$$

附：抽样分布

- **关系图**

$$(1)\ \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n});$$

$$(2)\ U = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1);$$

$$(3)\ \frac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1);$$

$\bar{X}$ 与 S^2 相互独立.

$$(4)\ T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

附2：一个模拟数据的实例

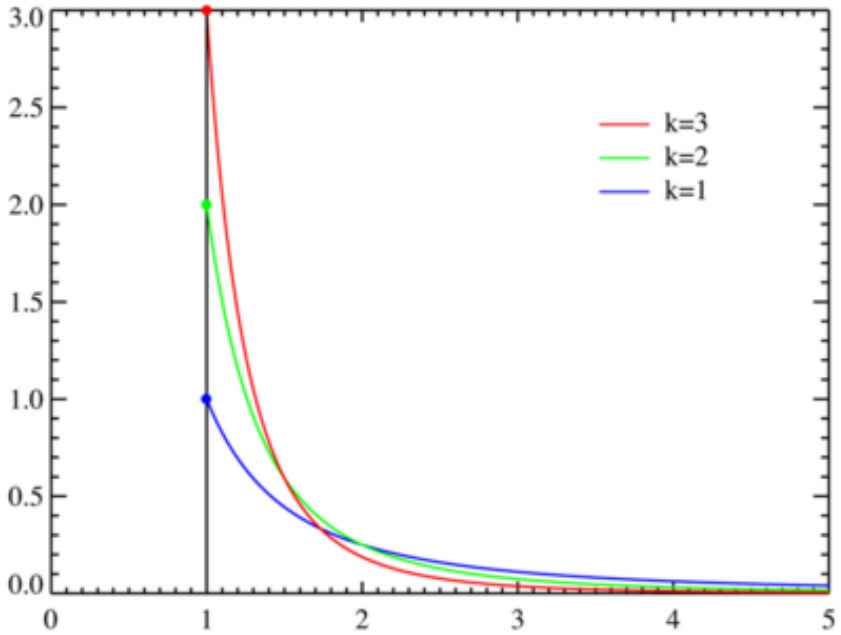
小象学院 ChinaHadoop.cn

➢ 帕累托分布

$$\mathrm{P}(X > x) = \left(\frac{x}{x_{\min}}\right)^{-k}$$

$$p(x) = \begin{cases} 0, & \text{if } x < x_{\min}; \\ \frac{k\, x_{\min}^{k}}{x^{k+1}}, & \text{if } x > x_{\min}. \end{cases}$$

期望值为 $\frac{x_{\min}\, k}{k-1}$。



图：帕累托分布（x_{min}=1）

被认为大致是帕累托分布的例子有：

- 在现代工业资本主义创造了大量中产阶级之前，财富在个人之间的分布。
- 甚至在现代工业资本主义创造了大量中产阶级之后，财富在个人之间的分布。
- 人类居住区的大小
- 对维基百科条目的访问
- 接近绝对零度时，爱因斯坦凝聚的团簇
- 在互联网流量中文件尺寸的分布
- 油田的石油储备数量
- 龙卷风带来的灾难的数量

附2：一个模拟数据的实例

小象学院 ChinaHadoop.cn

- **生成符合帕累托分布的数据， 且具有大于1000的点**
  - x=stats.pareto.rvs(b=1.2,loc=50,size=1000)
  - x.hist(bins=50)
  - plt.hist(x,bins=50)

2) Pareto Distributed Traffic: We generated flows with sizes drawn from a Pareto distribution with mean 50 KB and shape 1.2. This yields flow sizes that capture realistic data center workloads [1]. In these settings, L^{2}DCT improves over DCTCP at all loads as shown in Figure 4(c), with 99th percentile of FCT improved by up to 53%.

```
In [24]: p=1-stats.pareto.cdf(1000,b=1.2,loc=50)

In [25]: p
Out[25]: 0.00026713554171153842

In [26]: 1-stats.binom.cdf(1,2000,p)
Out[26]: 0.10074983217202338

In [27]: 1-stats.binom.cdf(1,20000,p)
Out[27]: 0.96967834903896866
```

小象学院
ChinaHadoop.cn

联系我们：

- 新浪微博：ChinaHadoop
- 微信公号：ChinaHadoop
- 网站：http://chinahadoop.cn
- 问答社区：http://wenda.ChinaHadoop.cn