第四课（第10-12课时）

理解样本数据

- 观察数据：继续使用pandas
- 抽样问题：对数据分布的检验
- 理论知识点3：假设检验

任务描述

- 进一步理解数据类型和数据结构：category
- 对数据聚合分组统计
- 检验数据的分布
- 理解假设检验

练习与问题

小象学院 ChinaHadoop.cn

- 1. 如何检验数据的抽样在某个维度是符合某种分布的？譬如，是否是正态分布，或，是否与总体的分布相同？
- 2. 用某两个维度构造的二维列联表，以及相应的可视化如柱状图，是否可以做出诸如“两组显著不同”这种结论？

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

数据处理的步骤

- **数据分析的步骤：**
  - 1. 获取数据
  - 2. 数据预处理
  - 3. 数据分析
  - 4. 数据挖掘
- **数据预处理：数据分析和挖掘的瓶颈**
  - 获取数据
  - 载入数据
  - 清洗数据：异常
  - 清洗数据：维度
  - 清洗数据：粒度
  - 缺失值；无效值；格式转换；命名变换；类型转换

载入数据：tips.csv

- **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
- **读入数据（假设文件在工作目录路径下）**
  - tips=pd.read_csv('tips.csv')
  - tips.describe()
  - 还有四列呢？
  - tips[['sex','smoker','day','time']].describe()
  - tips['sex'].value_counts()

```
In [6]: tips.describe()
Out[6]:
       total_bill         tip        size
count  244.000000  244.000000  244.000000
mean    19.785943    2.998279    2.569672
std      8.902412    1.383638    0.951100
min      3.070000    1.000000    1.000000
25%     13.347500    2.000000    2.000000
50%     17.795000    2.900000    2.000000
75%     24.127500    3.562500    3.000000
max     50.810000   10.000000    6.000000
```

Variable explorer

Name	Type	Size	
tips	DataFrame	(244, 7)	Column names: total_bill, tip, sex, smoker, day, time, size

DataFrame 索引总结

- .at, .iat, .loc, .iloc 和 .ix
  - tips['sex']
  - tips[0:3]

  - tips.at[1,'sex']
  - tips.iat[4,4]

  - tips.loc[:,['sex','size']]
  - tips.iloc[0,2]
  - tips.iloc[0:4,2:4]

DataFrame的合并

小象学院 ChinaHadoop.cn

- **按行合并**
  - tips1=tips[:100]
  - tips2=tips[100:]
  - tip12=pd.concat([tips1,tips2])
- **按列合并**
  - left = pd.DataFrame({'key': ['foo', 'foo'], 'lval': [1, 2]})
  - right = pd.DataFrame({'key': ['foo', 'foo'], 'rval': [4, 5]})
  - pd.merge(left, right, on='key')
- **Append**
  - s1=tips.iloc[4]
  - tips.append(s)
  - tips.append(s,ignore_index=True)

sort 和 groupby

- **排序**
  - tips.sort('tip')
  - tips.sort(['tip','total_bill'])
  - tips.sort(['sex','tip'],ascending=[True,False])
- **groupby**
  - tips.groupby('sex').sum()
  - tips.groupby(['sex','size']).sum()
  - tips.groupby(['sex','time']).mean()
- **apply**
  - tips[['total_bill','tip','size']].apply(lambda x: x.max()-x.min())

Categorical Data 类别型数据

- **类别型变量**
  - 有限的取值：如性别，星座，年级
  - 加减乘除没有意义
  - 分为有序的和无序的
    - 有序的如，改进程度；
    - 无序的如，性别；

```
In [32]: tips.dtypes
Out[32]:
total_bill      float64
tip             float64
sex            category
smoker           object
day              object
time             object
size              int64
dtype: object
```

- **类别型变量的产生**
  - s = pd.Series(["a","b","c","a"], dtype="category")
  - s = pd.Series(["a","b","c","a"])
  - s_cat = s.astype("category", categories=["b","c","d"], ordered=False)
    - 试试 ordered=True
  - tips['sex']=tips['sex'].astype('category')
  - tip.dtypes

Categorical Data 类别型数据

- 本质上是从逻辑上定义变量顺序
- 查看类别型数据
  - tips['sex'].describe()

```
In [41]: tips['sex'].describe()
Out[41]:
count      244
unique       2
top       Male
freq       157
Name: sex, dtype: object
```

- 是否有序
  - tips['sex'].cat.ordered
  - tips['sex']=tips['sex'].cat.set_categories(['Male','Female'], ordered=True)
  - s = pd.Series(pd.Categorical(["a","b","c","a"], ordered=False))
  - s.sort(inplace=True)
  - s = pd.Series(["a","b","c","a"]).astype('category', ordered=True)
  - s.sort(inplace=True)
  - s = s.cat.reorder_categories([2,3,1], ordered=True)

pandas练习

- **分组统计小费比例**
  - 按性别
  - 按时间（午餐、晚餐）
  - 按性别和时间

Scipy

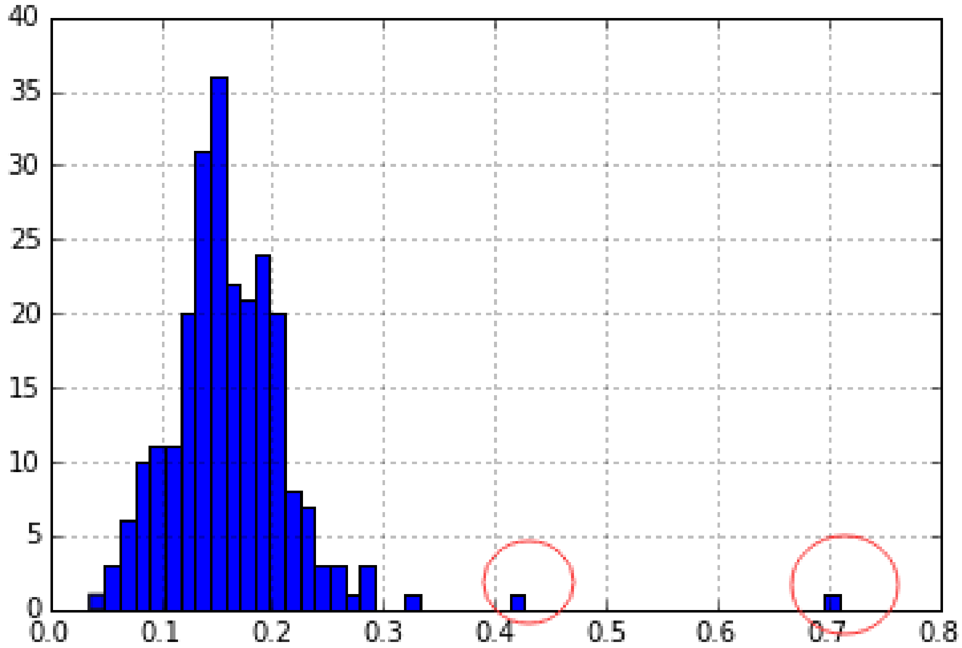
练习与问题

- 1. 如何检验数据的抽样在某个维度是符合某种分布的？譬如，是否是正态分布，或，是否与总体的分布相同？
- 2. 用某两个维度构造的二维列联表，以及相应的可视化如柱状图，是否可以做出诸如“两组显著不同”这种结论？

- **导入scipy中的统计库**
  - from scipy import stats

可视化分析

- **直方图**
  - tips1['tips_pct']=tips1['tip']/tips1['total_bill']
  - tips1['tips_pct'].hist(bins=50)



用数字特征检验

- 看分布的数字特征是否与理论的数字特征一致？
- 例：生成一个t分布序列
  - np.random.seed(282629734)
  - x = stats.t.rvs(10, size=1000)
- 计算其数字特征
  - print x.max(), x.min(),x.mean(),x.var()
  - 或：n, (smin, smax), sm, sv, ss, sk = stats.describe(x)
  - print n, (smin, smax), sm, sv, ss, sk
- 计算理论数字特征
  - m, v, s, k = stats.t.stats(10, moments='mvsk')
  - print m,v,s,k
- 似乎还不够。。。

更严格的检验

- **单一样本检验，是否为**
  - print 't-statistic = %6.3f pvalue = %6.4f' % stats.ttest_1samp(x, m)
  - “t-statistic = 0.391 pvalue = 0.6955”
    - 如何解读？

K-S test

- The Kolmogorov-Smirnov test
  - stats.kstest(x, 't', (10,))
  - print 'KS-statistic D = %6.3f pvalue = %6.4f' % stats.kstest(x, 't', (10,))
  - “KS-statistic D = 0.016 pvalue = 0.9606”
    - 如何解读？？
  - 是否符合正态分布？
  - stats.kstest(x, 'norm', (x.mean(),x.std()))
  - KS-statistic D = 0.032 pvalue = 0.2402
    - 如何解读？

卡方检验

小象学院 ChinaHadoop.cn

- quantiles = [0.0, 0.01, 0.05, 0.1, 1-0.10, 1-0.05, 1-0.01, 1.0]
- crit = stats.t.ppf(quantiles, 10)
- crit
- n_sample = x.size
- np.histogram(x, bins=crit)
- freqcount = np.histogram(x, bins=crit)[0]
- freqcount

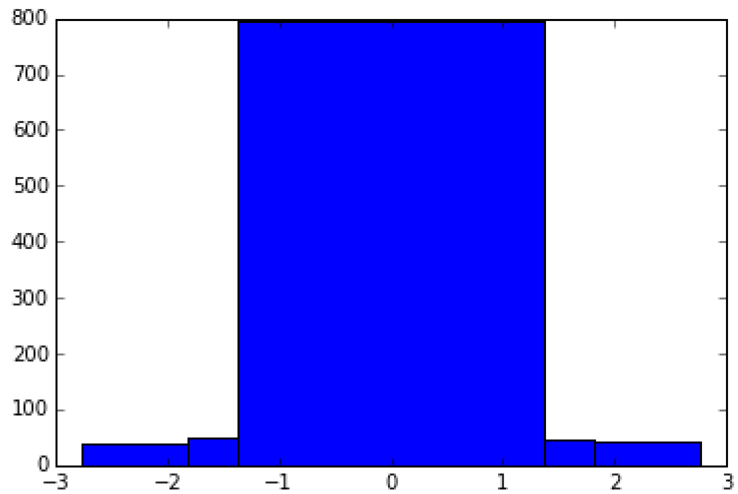
卡方检验

小象学院 ChinaHadoop.cn

- tprob = np.diff(quantiles)
- nprob = np.diff(stats.norm.cdf(crit))
- plt.hist(x,bins=crit)
- nprob = np.diff(stats.norm.cdf(crit))
- tch, tpval = stats.chisquare(freqcount, tprob*n_sample)
- nch, npval = stats.chisquare(freqcount, nprob*n_sample)
- stats.chisquare(freqcount, tprob*n_sample)
- freqcount
- tprob*n_sample

卡方检验

- print 'chisquare for t: chi2 = %6.2f pvalue = %6.4f' % (tch, tpval)
  - chisquare for t: chi2 = 2.30 pvalue = 0.8901
- print 'chisquare for normal: chi2 = %6.2f pvalue = %6.4f' % (nch, npval)
  - chisquare for normal: chi2 = 64.60 pvalue = 0.0000



基于样本估计的数字特征进行卡方检验

小象学院 ChinaHadoop.cn

- tdof, tloc, tscale = stats.t.fit(x)
- nloc, nscale = stats.norm.fit(x)
- tprob = np.diff(stats.t.cdf(crit, tdof, loc=tloc, scale=tscale))
- nprob = np.diff(stats.norm.cdf(crit, loc=nloc, scale=nscale))
- tch, tpval = stats.chisquare(freqcount, tprob*n_sample)
- nch, npval =stats.chisquare(freqcount, nprob*n_sample)

- 在5% level，拒绝正态分布

什么是p值？---假设检验

- 假设检验是指施加于一个或多个总体的概率分布或参数的假设. 所作假设可以是正确的, 也可以是错误的.
- 为判断所作的假设是否正确，从总体中抽取样本, 根据样本的取值, 按一定原则进行检验，然后作出接受或拒绝所作假设的决定.

参数假设检验是指总体X的分布函数$F(x;\theta)$的类型已知，参数θ未知，首先对未知参数θ提出假设："θ_0为其真值"然后由抽取的样本所提供的信息对假设的正确性进行判断的过程.

非参数假设检验是指总体X的分布函数$F(x)$未知，首先假定其分布函数为某指定函数$F_0(x)$提出假设,然后根据样本信息来检验这个假设，最后做出拒绝或接受的判断.

假设检验

小象学院 ChinaHadoop.cn

➢ 例

例8.1.1 某车间用一台包装机包装食盐。包装的每袋食盐的重量是一个随机变量且服从正态分布，当机器正常时，其均值为0.5公斤，标准差为0.015公斤。某日开工后为检验包装机是否正常，随机抽取9袋，称得净重量为（公斤）：0.497，0.506，0.518，0.524，0.498，0.511，0.520，0.515，0.512。假设标准差保持不变，问包装机是否正常？

作假设：$\mu = 0.5$

一般把不轻易否定的命题作为原假设

我们要做的是：根据样本检验所做的假设是否为真

假设检验

- **假设检验的基本思想**
- **小概率原理：** **概率很小的事件在一次试验中几乎不可能发生. 如果发生了，就认为不合理，应否定. 即假设不成立.**

在假设检验问题中，把有关总体未知分布的假设称为统计假设，简称假设

把待检验的假设称为原假设或零假设，记为 H_0

与之对立的假设称为备择假设或对立假设，记为 H_1

一个假设检验问题通常简记为 $H_0 \leftrightarrow H_1$

比如：H_0：$\mu=18.2 \leftrightarrow H_1 : \mu \neq 18.2$

注：在处理问题时，应把着重考察且便于处理的问题作为 H_0

假设检验

小象学院 ChinaHadoop.cn

- 一个例子

例8.1.1 某车间用一台包装机包装食盐。包装的每袋食盐的重量是一个随机变量且服从正态分布，当机器正常时，其均值为0.5公斤，标准差为0.015公斤。某日开工后为检验包装机是否正常，随机抽取9袋，称得净重量为（公斤）：0.497，0.506，0.518，0.524，0.498，0.511，0.520，0.515，0.512。假设标准差保持不变，问包装机是否正常？

待检验假设：$\mu = 0.5$

如何判断它的正确性

假设检验

小象学院 ChinaHadoop.cn

- **一个例子**

如果$H_0: \mu = \mu_0$为真，$\bar{x}$偏离μ_0仅仅是由于随机误差的原因，那么$\bar{x}$会以很大概率落在μ_0附近一定的范围内，而远离μ_0的概率会很小，即$|\bar{x} - \mu_0|$一般不应太大.

当H_0成立时，$\dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$，

衡量$|\bar{X} - \mu_0|$的大小就归结衡量$\dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$的大小.

考虑一个小的正数λ，$\left|\dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}\right|$不应超出$\lambda$，即

$$\left\{\left|\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}\right| > \lambda\right\}$$是一个小概率事件.

假设检验

➢ **一个例子**

在假设$H_0: \mu = \mu_0$成立时，有统计量

$$U = \frac{\bar{X} - 0.5}{0.015}\sqrt{9} \sim N(0,1)$$

对于小概率$\alpha = 0.05$, 查表可得$u_{\alpha/2} = 1.96$

$$P\{|U| > 1.96\} = P\left\{\left|\frac{\bar{X} - 0.5}{0.015/\sqrt{9}}\right| > 1.96\right\} = 0.05$$

小概率事件

假设检验

- 一个例子

由题意$\overline{x} = 0.511$

$$|U| = \left|\frac{0.511 - 0.5}{0.015/\sqrt{9}}\right| = 2.24 > 1.96$$

小概率事件发生了！

假设检验

- **一个例子**

如果抽样的$\bar{x}$实际上远离μ_0而使得小概率事件发生了，也就是说：不太可能发生的事情发生了，我们自然有充足的理由否定假设$H_0:\mu=\mu_0$的正确性。

相反，如果样本值$\bar{x}$没有远离μ_0从而没有使得小概率事件发生，也就是说：$\bar{x}$偏离μ_0可以认为是由于样本随机性造成的合理偏离，我们没有从样本信息中找到足够的证据来否定假设$H_0:\mu=\mu_0$的正确性，于是，不能拒绝假设H_0

假设检验

- **概念**

小概率α---假设检验的显著性水平

小概率事件-------由样本描述的概率不超过α的事件

拒绝域(否定域)----拒绝H_0的样本值的取值区域.

$$C=\{(x_1,x_2,\cdots,x_n):\ |U|>u_{\alpha/2}\}$$

接受域----接受H_0的样本值的取值区域.

$$\bar{C}=\{(x_1,x_2,\cdots,x_n):\ |U|\le u_{\alpha/2}\}$$

双边(侧)检验、单(侧)边检验

假设检验两类错误

1、第一类错误 -----弃真错误

$$P(\text{拒绝}H_0 \mid H_0\text{真}) \le \alpha$$

注：可用显著性水平的大小来控制犯第一类错误的概率的大小

那么，是否显著性水平越小，假设检验的准确性就越高呢？

事实上不然，因为，一般来说，当样本容量给定时，在降低显著性水平的同时，往往会增大犯第二类错误的可能性。

假设检验两类错误

2、第二类错误------纳伪错误

$$\beta = P(\text{接受}H_0 \mid H_0\text{不真})$$

3、关系

当样本容量n固定时，α与β不能同时都控制得很小。

若要使犯两类错误的概率都减小,除非增加样本容量.

假设检验

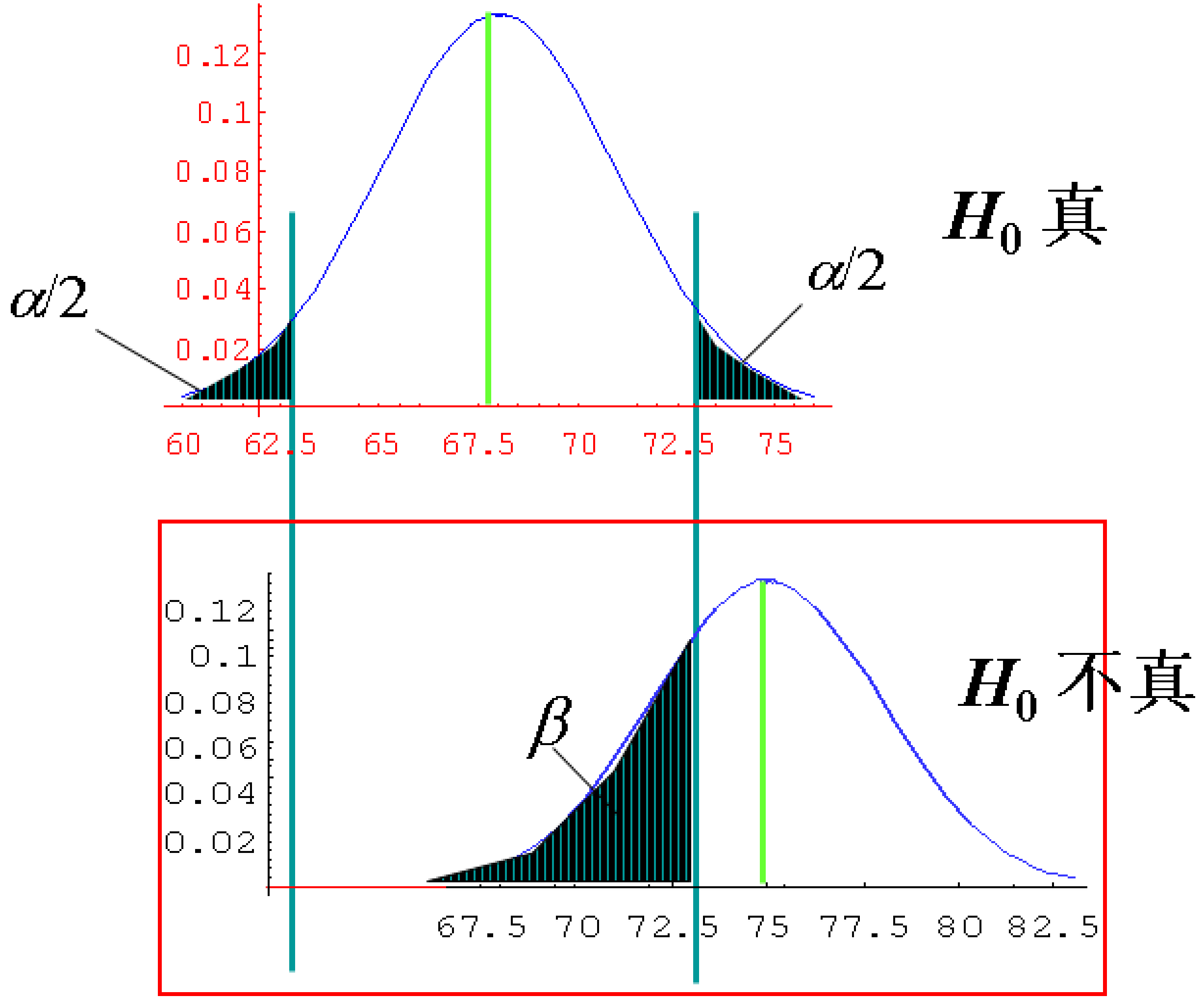
小象学院 ChinaHadoop.cn

- **统计假设检验的四种状况**

		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

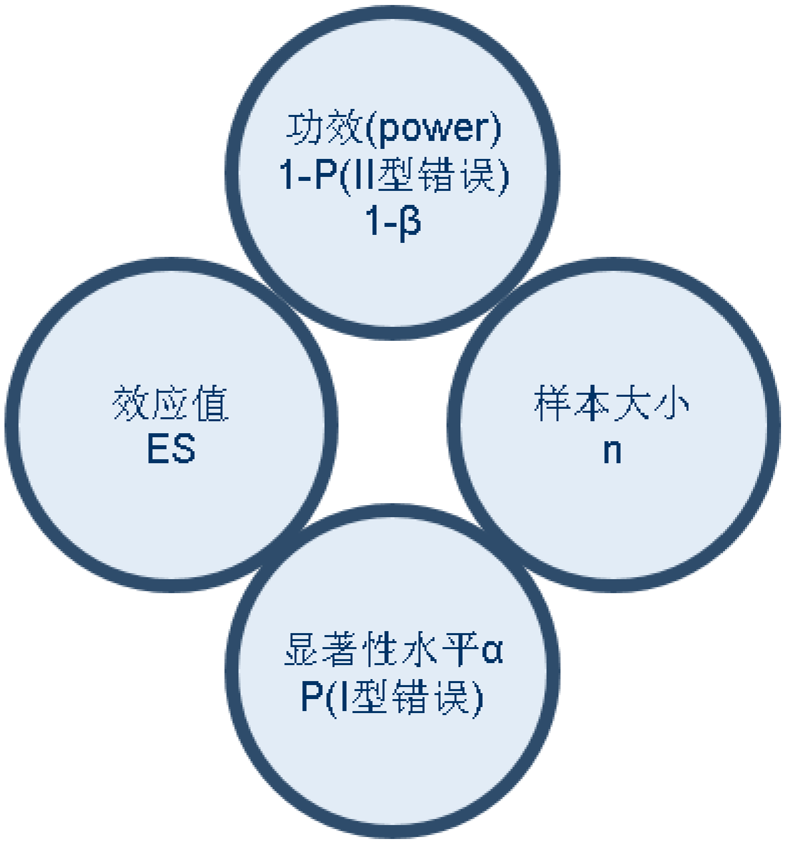
假设检验两类错误

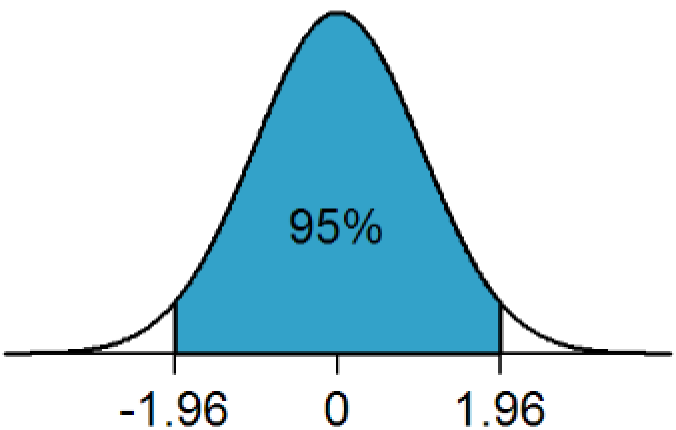
小象学院 ChinaHadoop.cn



假设检验

- **相关课题：功效分析**



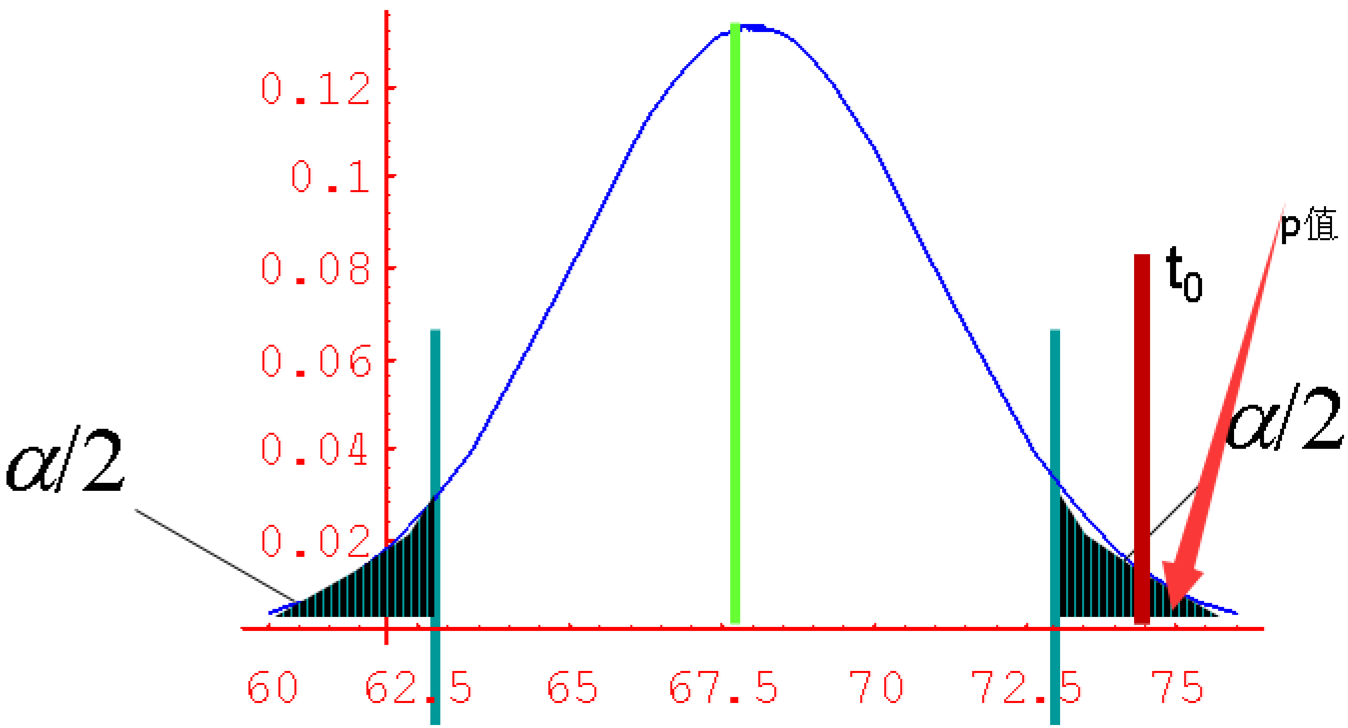


假设检验：什么是p值

计算p值

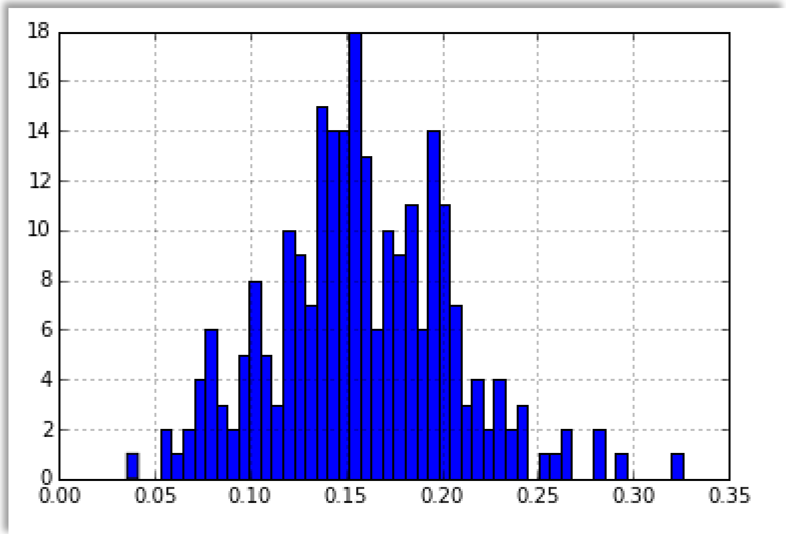
$$p = P\{|T_0| > t_0\}$$

比较α值与p值并作结论



正态分布检验

- stats.normaltest(x)
- stats.normaltest(tips['tip'])
- tips['tip_pct']=tips['tip']/tips['total_bill']
- tips.tip_pct.hist()
- stats.normaltest(tips['tip_pct'])
  - 异常点引起的？
- tips[tips['tip_pct']>0.4]
- tips1=tips.drop([172,178])
- stats.normaltest(tips1['tip_pct'])



比较两个样本

- **比较两组均值**
  - rvs1 = stats.norm.rvs(loc=5, scale=10, size=500)
  - rvs2 = stats.norm.rvs(loc=5, scale=10, size=500)
  - stats.ttest_ind(rvs1, rvs2)

  - rvs3 = stats.norm.rvs(loc=8, scale=10, size=500)
  - stats.ttest_ind(rvs1, rvs3)

- **比较两组分布**
  - Kolmogorov-Smirnov 双样本检测 ks_2samp
  - stats.ks_2samp(rvs1, rvs2)
  - stats.ks_2samp(rvs1, rvs3)

复习：样本分布

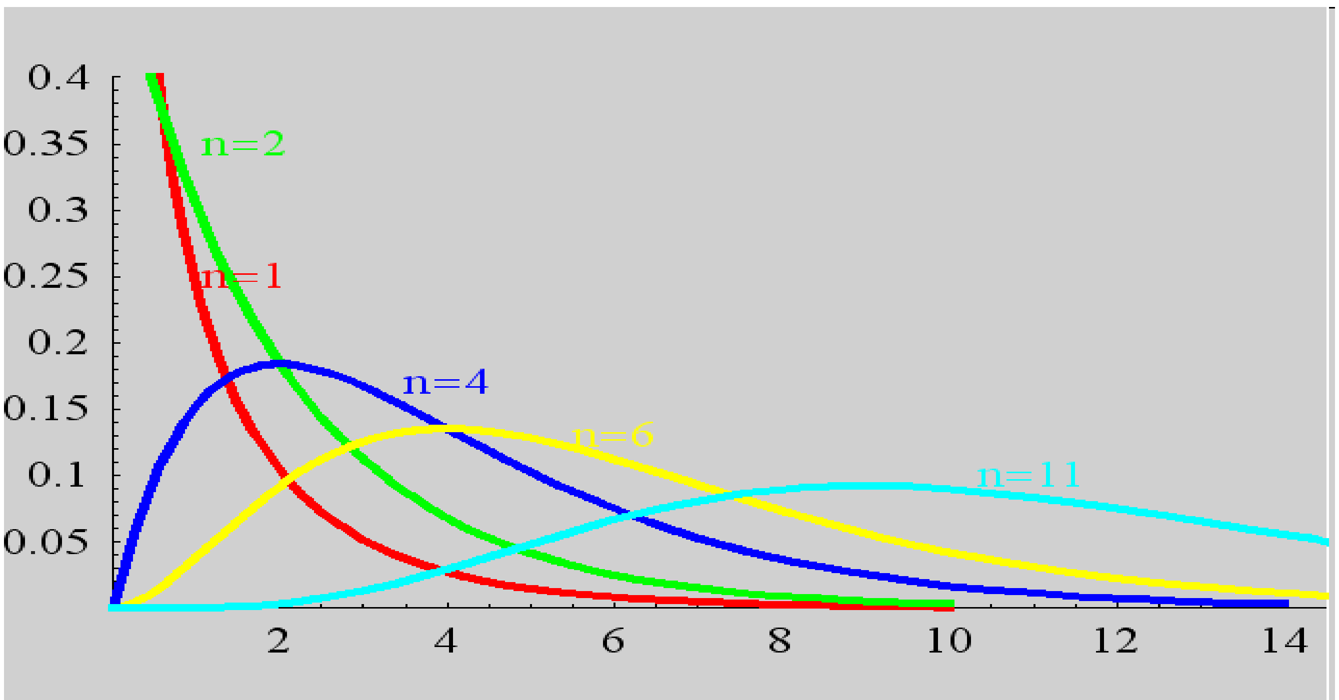
- 正态分布
- (student)t分布
- 卡方分布
- F分布
- 扩展阅读：

小象学院 ChinaHadoop.cn

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

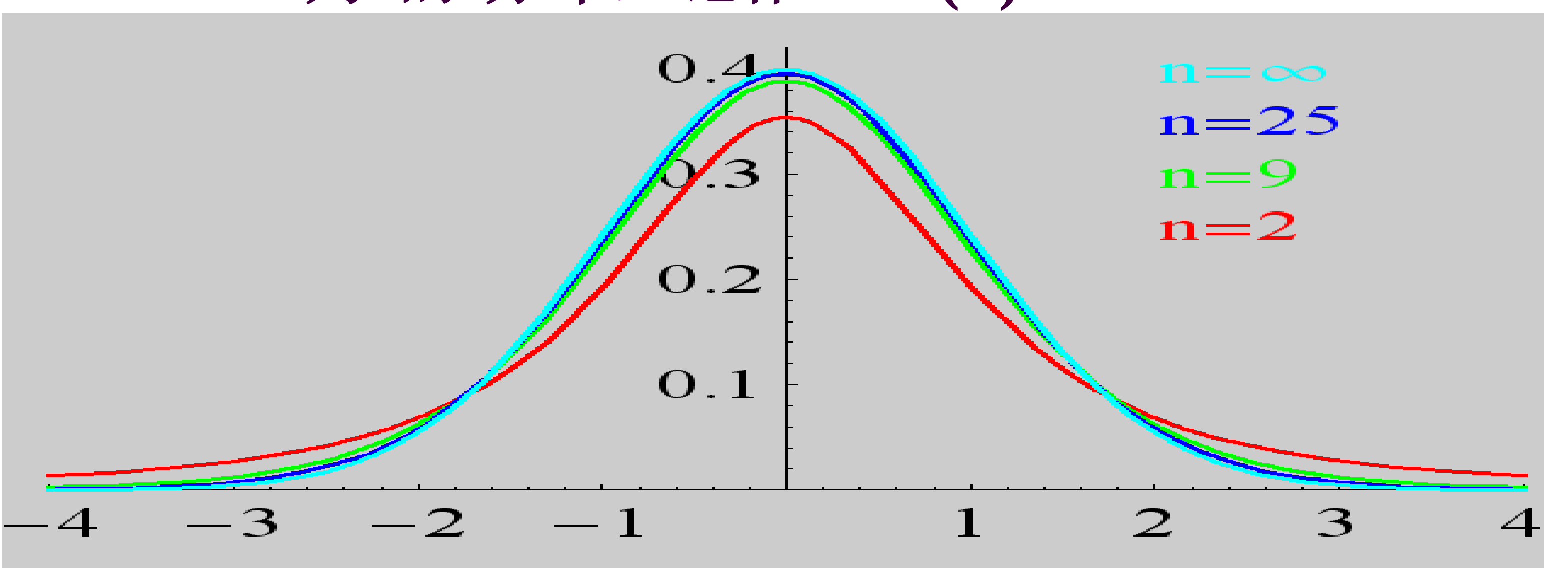


抽样分布

小象学院 ChinaHadoop.cn

➢ ***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$T \sim t(n)$.



抽样分布

- t 分布

设总体 $X \sim N(\mu,\sigma^2)$，$X_1, X_1, \cdots, X_n$（$n \ge 2$）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.

- ***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 $X^{-1}\sim F(n,m)$.

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

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

抽样分布

- 关系图

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

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

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

$\bar{X}$ 与 S^2 相互独立.

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

练习

- **根据tips数据，设计并完成以下检验**
- **周四到周日，每日用餐人数相等，使用何种检验？**
  - A. normal， B. t-test，C. 卡方，D. K-S
- **男女性别给的小费比例不同**
  - A. normal， B. t-test，C. 卡方，D. K-S
- **性别每日分布相同**
  - A. normal， B. t-test，C. 卡方，D. K-S

联系我们：

- 新浪微博：ChinaHadoop
- 微信公号：ChinaHadoop
- 网站：http://chinahadoop.cn