统计描述　＋　统计推断

<Statistics>统计学

统计学定义为收集数据，分析数据，由数据得出结论而组成的概念，原则和方法

统计可被定义为在随机性中寻找规律性

经验变量：特征，字段，属性

理论变量：经验变量→变换，如z, t, gamma, F变量

抽样误差：如果研究被再做一遍，结果未必会和上次一模一样。抽样误差告诉我们，样本离总体的实际值可能有多远

抽样误差的大小依赖于得到样本的方式和样本中包含的观测的个数。

在公布任何一次抽样调查的结果时都应说明抽样误差的大小，不管是比例，均值还是其它形式

标准得分：(观察值-均值)/标准差

不同的变量一般有不同的均值和标准差，统计上，均值和标准差不同时，一个变量的值不能与另一变量的值相比较。将变量值转换化标准得分，从而一个变量的任何值都可以和任何其它变量的值相比较，因为我们知道任何一个得分与均值的相对距离

mean-2\*std mean-std mean mean+std mean-2\*std

-2 -1 0 1 2

任何变量的标准得分的值大部分落在[-2, 2]，2/3的标准得分值落在[-1, 1]。如果一个观察值的标准得分>2 or <-2，那么这个观察值就不寻常的大或小。标准得分常称为t值

统计推断：从样本数据得出与总体参数值有关的结论　＝　估计＋假设检验

尽管样本中的信息不完全，而且来自样本的结果一般者不等于总体真值，为了弥补样本结果的不准确，研究者们计算抽样误差－－这个数能使95%的抽样结果都位于由总体真值加减样本误差所得到的区间内

点估计：是一个用来估计总体参数的数，一般用样本统计量作为总体参数的点估计

最大似然法，Bayes估计法

注意：特定数据训练得到的估计方法的准确性不能通过相同的数据进行测试，（举个例子，考试题目如果与课堂习题完全相同，对知识的理解效果就得不到检验，因为仅靠死记硬背就能得到最佳答案。所谓的实力，是解决之前未曾遇到过的问题的能力）这涉及到模型选择的方法：一般采用交叉检验法

置信区间：包含总体参数真值的区间　＝　点估计＋抽样误差

无法回答某个特定的置信区间是否包含总体真值，我们知道的仅仅是－－在多次抽样中有95%的样本得到的区间包含真值，95%称为置信水平

置信区间是一个随机区间，因为置信区间会因样本的不同而不同，而且不是所有的区间都包含参数真值。

一次抽样无法回答，因为每次抽样的样本集不同，导致点估计和区间估计均不同，但无数次抽样，得到的无数个点估计和区间估计，

无数个点估计的平均为总体真值；

无数个区间估计中，至少有95%的区间估计中包含真值，则95%称为置信水平

换句话，95%置信水平是指多次抽样中有95%的置信区间包含未知的总体参数值，至于在一次抽样得到的置信区间是包含总体参数的众多区间中的一员呢，还是属于个别不含参数值的区间就永远不得而知了。

置信区间特点：

大的样本产生较短的置信区间

低的置信水平产生较短的区间

按照惯例，抽样误差默认是基于95%置信水平的，所有好的调查都应该提供抽样误差，这样读者自己就可以构造置信区间了

多个总体，抽取各自的样本，若样本均值有差异，能否推出总体均值有差异？若有差异，则称为统计显著

因为样本均值的差异，有可能仅仅归于随机性

零假设：假设的内容总是没有差异，或没有改变，或变量间没有关系等等

零假设所提问题的答案是以样本数据为基础的，值得强调的是零假设总是一个与总体参数有关的问题，关于样本统计量如样本均值或方差的零假设是没有意义的，因为样本统计量是已知的

零假设：两人总体均值无差异

假设检验机制：

将样本均值之差1.3，标准差及样本数　→ 　标准得分(t变量)　→ 查t分布表，得出样本均值之差>1.3的概率，即p-value, → 若p-value < 0.05

两种解释：

* 零假设是正确的，观测到的数据恰好是不常发生的那一类；
* 数据倒是常见的那一类，只是零假设是错的

若零假设为真，那么样本均值差异1.3应很常见，而实际计算出基本样本均值差异1.3的概率小于0.05, 说明不常见，那么选择第二种解释，即认为导致这个小概率的假设－－两总体均值无差异是错的，即应认为两人总体均值差异不为0.

总结：首先要对我们研究的事物作出某种假设，然后收集数据，并在假设的基础上计算得到该数据的概率，如果这个概率非常小，如小于0.05，则一开始的假设就是错误的

假设检验只考虑到一个可能值，置信区间给了我们一个参数值的可能范围。即使当我们拒绝了零假设，并得到均值之差不为零的结论之后，紧接着的一个问题就是该差异有多大，这个问题是由置信区间来负责回答

变量间关系

许多变量可能有统计关系但没有因果关系

比如冰淇淋销量与车祸中儿童受伤数，只是统计关系，背后的第三个变量，温度则是与这个变量有因果关系

用一个变量的值预测另一个变量的值时，它们之间不必非得有因果关系

自变量：先发生的变量

因变量：受影响的变量

两个分类变量的相关分析　－－卡方分析

列联表：描述两个分类变量分布的频率表　contingency table

若自左上角至右下角的对角线上有更多的观测，这种频率结构表明在这些数据中这两个变量是相关的

若两列的百分比分布不同，则两变量存在着关系

若相关系数<0.3，认为关系比较弱; 若[0.3 0.7],则关系适中；若>0.7,则关系强

相关系数计算见公式

两个数值型变量的回归分析和相关分析

离群点对相关系数影响大

列差变量：除自变量外其他所有变量对因变量的效应 = sum(yi - a-bxi)2 观测值－估计值

总的平方和：度量所有的变量对因变量的效应　＝ sum(yi – u)2 观测值－均值

回归平方和：度量自变量对因变量的效应 ＝sum( a+bxi - u)2 估计值－均值

r = sqrt(回归平方和/总的平方和)

相关系数的平方　＝　因变量取值变化的效应可归于自变量的比例

分类自变量和数值因变量的方差分析

若分类自变量只有两种取值（如西海岸和东海岸），那么可以定义虚拟变量，西海岸映射为0, 东海岸映射为1，从而做回归分析

若分类自变量有多种取值，采用方差分析anova

方差分析：用来对比因变量在不同组中的平均值的统计方法

常用盒子图来对比各组中因变量的不同

假定自变量区域，因变量暴力率

方差分析：先根据自变量（地区）分组，再求出每一组的因变量的平均值，我们感兴趣的是因变量的均值在各组之间是否有差异，需要借助方差

自变量平方和：　sum(组均值－总均值)2 度量地区变量的效应

残差平方和：sum(观测－组均值)2

总平方和：sum(观测－总均值)2 = 自变量平方和+残差平方和

r = sqrt(自变量平方和/总平方和)

地区变量与暴力犯罪率变化的R2有关联，而残差变量与剩下的1-R2有关联

多元分析

在多元分析中，残差变量对因变量的影响减少了，这是因为我们将所有自变量的作用同时从残差变量中分离出来，而非逐一分离

统计方法的选择总是决定于所考虑的变量的性质

自变量 因变量 方法

分类型 分类型 偏phi法

数值，哑元 数值型 多元回归

回归模型无截距项，有m个特征，设置

偏phi法：

控制其余变量，计算单变量和因变量的phi

然后取各组中的phi的均值

回归，分类器

若自变量是数值型和分类型，对分类型变量引入哑元

若某分类变量level=5, 则引入5个哑元变量，每个哑元变量取值0/1，每个二进制特征对应分类数据的一个等级

categorical\_feature = pd.Series(['sunny', 'cloudy', 'snowy', 'rainy', 'foggy'])

dummies = pd.get\_dummies(categorical\_feature)

如何生成概率分布？

伪随机序列：由随机的种子确定的长序列，只有种子是随机的

真随机：长序列的每个元素都是随机的

代码库一般提供[0 1]范围内的均匀分布函数，对生成的随机序列进行变换，使其遵从特定的分布

0-1分布： y = 正面朝上　if x < 0.5

均匀分布：y = (b-a)\*x + a

正态分布：y = 0.5\*exp(-x\*x/2) and z = std\*y+mean

多元正态分布：N(mean\_vec, cov\_mat)

随机生成变量(x1, x2, ...xn), 从而求f(x1, x2, ...xn)

采用scipy.stats工具包

from scipy import stats

# 二项分布binomial distribution

n次事件中，发生概率为0.5, 恰好发生x次的概率

#计算概率值p(x=650), p(x=651), ...p(x=1200), n=1200, p=0.5

stats.binom.pmf(range(650, 1201), 1200, 0.5)

#生成随机变量x, 服从二项分布n=1200, p=0.5

data = stats.binom.rvs(n=1200, p=0.5, size=1000)

plt.hist(data, bins=24, normed=True) #可视化随机变量pmf

# 泊松分布 Poisson distribution

某路口事件发生的比率是ramda(每天发生次数), 求一段固定时间间隔内(这里是1天）发生x次事件的概率

stats.poisson.pmf(range(0, 10), 2)

#生成随机变量x, 服从泊松分布ramda=2

data = stats.poisson.rvs(mu=2, size=1000)

# 正态分布

stats.norm.pdf(np.arange(-5, 5, 0.1), loc=0, scale=1)

data = stats.norm.rvs(loc=0, scale=1, size=1000)

数据分析

类别型变量~类别型变量

独立检验: 卡方检验 （p值：独立概率,若p值小，则拒绝二者独立）

Library(vcd)

Mytable <- xtabs(~Treatment+Improved, data=Arthritis)

Chisq.test(mytable)

相关性度量:相关性强弱系数Phi

Assocstats(mytable)

数值型变量~二元类别型变量 （即样本分成两个组）

用于检验两个总体的均值是否相等？（若p值小，拒绝两组相同）

两组样本独立的t检验，且从正态总体中抽样

t.test(数值变量~二元变量,data=..)

可视化：盒子图boxplot(数值型~类别型)

两组样本不独立的t检验，组间差异呈正态分布

数值型变量~多元类别型变量(即样本分成多个组)

方差分析：名字令人误解，更适合的名字是均值分析：先根据自变量分组，再求出每一组因变量的均值，目的是讨论各组的均值是否不同。但这里判断均值之间是否有差异是借助于方差.

Fit <- aov(数值型变量~类别型变量)

可视化：带有置信区间的组均值图形

Library(gplots)

Plotmeans(数值型变量~类型别变量)

boxplot

机器学习

SVM

两类分类通常用一个实数函数：

线性分类: f(x)是线性函数，

非线性分类：非线性可分时，把样本x映射到某个高维特征空间，在高维特征空间使用线性学习器。

数据挖掘与临床诊断

[**数据挖掘与临床疾病诊断**](http://blog.csdn.net/hitheu/article/details/8037336)

数据挖掘（Data Mining）通过分析每个数据，从大量数据中寻找规律的技术。主要有数据准备、规律寻找和规律表示三个步骤。数据准备是从相关数据源中选取合适数据并整 合成用于数据挖掘的数据集；规律寻找是用某种方法将数据集合所含的规律找出来；规律表示是将找出的规律进行可视化。数据挖掘用到了统计学的抽样、估计、假 设检验的思想，说到底数据挖掘得出的规律就是分析从用户当前数据得到满足某一概率的潜在数据。通过数据挖掘技术我们可以做以下几种事情：分类（从数据中选 出已经分好类的训练集，在该训练集上运用数据挖掘分类的技术，建立分类模型，对于没有分类的数据进行分类。 ）、估计（与分类类似，不同之处在于，分类描述的是离散型变量的输出，而估计处理连续值的输出；分类的类别是确定数目的，估计的量是不确定的）、预测（预 测是通过分类或估值起作用的，通过分类或估值得出模型，该模型用于对未知变量的预言。预言其实没有必要分为一个单独的类。其目的是对未来未知变量的预测， 这种预测需要经过一定时间，才知道预言准确性是多少）、相关性分组（Affinity grouping or association rules决定哪些事情将一起发生）、聚类（对记录分组，把相似的记录在一个聚集里。聚类和分类的区别是聚集不依赖于预先定义好的类，不需要训练集。）、 描述和可视化（Description and Visualization是对挖掘结果用用户能够理解的方式展示）、复杂数据类型挖掘（主要指图形图形、视频音频等复杂数据的挖掘）。  
        关联规则是数据中所蕴含的一个重要规律，对关联规则的挖掘的目标是在数据项目中找出所有的并发关系。关联规则的挖掘广泛应用于社交网络中的好友推荐、购物 网站中的商品推荐、舆情分析系统中的與情走向预测、汉语输入法的智能关联等。一个经典的关联规则挖掘应用的例子零售品销售大王沃尔玛的“尿布与啤酒”：沃 尔玛公司拥有世界上最大的数据仓库系统，为了准确获得用户的购买习惯，沃尔玛对用户的购物行为进行了分析，想知道顾客经常一起购买的商品有哪些。发现了与 尿布一起购买的最多的商品是啤酒。于是沃尔玛零售店将其放在一起销售，结果发现它们的销量双双增加了。按常规思维，尿布与啤酒风马牛不相及，若不是借助数 据挖掘技术对大量交易数据进行挖掘分析，沃尔玛是不可能发现数据内在这一有价值的规律的。  
        在医院临床实践的疾病诊治过程中，一般都是通过患者的自述和医生以及仪器检查（主要是血常规、CT、X光、核磁共振等）得出的数据 ，然后医生根据以上数据加之自己的经验判断最有可能患的疾病类型，按照该种疾病的治疗方案进行治疗，然后复查治疗结果并对治疗方案进行修正，直到治愈出院 为止。对于以上的治疗过程我们可以建立一个数学模型来模拟：设患者自述与医生仪器检查得出的数据集合为A，疾病类型的集合为B，治疗方案的集合为C，整个 治疗过程用形式化描述其实就是由A推出潜在的B，由B得出C，复查与改进治疗方案也是这个过程。不同的疾病可能有相同的症状，医生则需要根据自己的临床经 验得出最有可能的疾病；当然同一疾病可以有不同的治疗方案，医生需要根据病人的身体状况、经济条件选择合适的治疗方案。由于在医疗疾病诊治过程中要求具有 较高的精确度，所有我们可将最小支持度min\_sup和最小置信度min\_conf设置较高，比如90%以上。为了实现在辅助的疾病诊治系统，医院需要建 立一个数据库以包含所有的病历，主要的记录内容：临床症状表现、体温、血常规、x光、ct等仪器检查结果、疾病名称、可行的急症处置方案、后续治疗方案、 治疗反馈信息等。病人来了之后，将所有检查得出的信息输入计算机，通过Apriori算法进行处理得出{临床病症表现}-->{疾病、处置方案}这 样一个关联规则，提供给医生参考，医生在根据实际的情况作出相应的选择。这个系统运行过程中可以是一个自学习的过程，通过不断添加病例情况、合并相同病 例，可以不断的丰富这个样例库，进而可以使得得出的这个关联规则可行度提高，辅助的诊治结果不断精确。这样的一个系统具有一定的应用前景，因为在实际的诊 治过程中，医生需要通过有限的数据在短时间内得出一个可行的治疗方案，系统的推荐结果可以给医生提供一些思路，这样的结果是通过大量的临床实践与众多专家 经验总结的结果。但在系统的建立过程中可能会遇到下列一些问题：首先病人对同一症状的表述可能由由不同的语言，医生需要进行整理用专业化术语表达；其次不 同医生对同一病症有不同的描述；再次如何从多种可能的疾病或则治疗方案中推荐一种最合适的方案；最后就是个体差异，也就是说同样的疾病在不同的人生上有不 尽相同的临床表现，疾病的临床表现没有一个绝对的界限，比如界定咳嗽严重程度等。现阶段人工智能领域的自然语言处理还不能做到精确分析人类语言的程度，这 为系统的实现带来了些难度，个体的差异为系统诊治结果带来了一份不确定性。这些问题的解决不仅需要从理论上加以深化研究，而且需要对模型进一步细化。

自然语言处理

概率: 从随机试验中的事件到实数域的映射函数，用以表示事件发生的可能性.

估计概率值：相同的情况下重复试验N次，当N足够大时，用相对频率近似概率值

条件概率：如果A,B是样本空间的两个事件，在已知事件B发生的情况下，事件A的概率。

条件概率计算：（贝叶期法则）

随机变量：一个随机试验可能有多种不同的结果，到底会出现哪一种，存在一定的概率。简单地说，就是试验结果的函数 (也称为随机变量的概率分布)

在自然语言处理中，一般以句子为处理单位，为了简化问题的复杂性，通常假设一个句子的出现独立于它前面的其他语句，句子的概率近似地被认为符合二项式分布。

联合概率分布和条件概率分布

贝叶斯决策理论

信息论

信息熵：信息量的量化，即描述一个随机变量的不确定性的数量。物理意义：表示信源X每发一个符号（不论发什么符号）所提供的平均信息量。熵越大，不确定性越大（即信息量越大），那么正确估计其值的可能性越小。

在已知部分知识的前提下，关于未知分布最合理的推断应该是符合已知知识最不确定或最大随机的推断.

最大熵的应用：选择模型

自然语言处理中，通常的做法是，根据已知的样本设计特征函数，假设存在k个特征函数

，它们都在建模过程中对输出有影响，那么，所建立的模型应满足所有这些特征的约束，即所建立的模型p应该属于这k个特征函数约束下所产生的所有模型的集合C.使熵H(p)值最大的模型用来推断某种语言现象存在的可能性，或者作为进行某种处理操作的可靠性依据。

联合熵：描述一对随机变量平均所需要的信息量。

长度为n信息，每一个字符的熵:

互信息：反映的是在知道Y的值以后，X的不确定性的减少量。可以理解为Y的值透露了多少关于X的信息量。互信息体现了两变量之间的依赖程度：如果I(X,Y)>>0,表示X与Y是高度相关的话，如果I(X,Y)=0，表明X和Y是相互独立的.

相对熵:也称KL距离，是衡量相同事件空间里两个概率分布相对差距的测度.

交叉熵：用来衡量估计模型与真实概率分布之间差异情况的。等于真实概率分布的熵+估计模型概率分布和真实概率分布之间的相对熵。在设计模型时，目的是使交叉熵最小，从而使模型最接近真实的概率分布. 设计语言模型时，一般用困惑度来代替交叉熵衡量语言模型的好坏。困惑度是交叉熵的指数函数.

噪声信道模型

信息熵可以定量地估计信息源每发送一个符号所提供的平均信息量。

一般情况下，在信号传输的过程中都要进行双重性处理：一方面要对编码进行压缩，尽量消除所有的冗余； 另一方面又要通过增加一定的可控冗余以保障输入信号经地噪声信道传输以后可以很好地恢复原状。

在自然语言处理中，不需要进行编码，一种自然语言的句子可以视为已编码的符号序列，但需要进行解码，使观察到的输出序列更接近于输入。

模拟信道模型，在自然语言处理中，很多问题都可以归结为在给定输出O（可能含有误传信息）的情况下，如何从所有可能的输入I中求解最有可能的那个，即求出使p(I|O)最大的I所有为输入I。用贝叶斯公式求解：

举例：如何想把一个法语句子f翻译成英语e,那么相应的翻译信道模型就是假定法语句子f作为信道模型的输出，它原本就是一个英语句子e，但通过噪声通道传输时被改变成了法语句子f.那么现在需要做的就是如何根据概率p(e)和p(f|e)的计算求出最接近原始英文句子e的解