统计学三大分布与检验

Li Junli
Mar 31 2019

理论基础——大数与中心极限定理

大数定理—— 当样本数量足够大时，这些样本的均值无限接近总体的期望

中心极限定理—— 无论总体服从什么分布，样本量足够大时，样本均值近似 *~N(u, σ^2/N)*

T分布与T检验

T分布

设 *X~N(0, 1), Y~χ 2(n)*，且X与Y互相独立

$$随机变量t=\frac{X}{\sqrt{Y/n}}服从自由度为n的t分布，记做t(n)$$

$$E(X)=0(n>1),\ D(X)=\frac{n}{n-2}(n>2),\ n很大时，近似正态分布$$

T检验

T检验是用t分布理论来推论差异发生的概率，从而比较两个平均数的差异是否显著。适用于小样本，总体标准差未知的情况。t检验可分为单总体(单样本)检验和双总体(两样本)检验，以及配对样本检验

Python模块—— scipy.stats

单样本T检验

前提：正态分布或近似正态分布

单样本t检验是检验样本平均数与已知总体平均数的差异是否显著，即检验样本是否来源于已知总体

$$t(n-1)=\frac{\overline{X}-u}{s/\sqrt{n-1}}$$

$$s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}},\ \overline{x}是样本平均数$$

在零假设：*μ=μ0*为真的条件下服从自由度为*n−1*的t分布

两样本T检验

前提：两样本独立且他们的总体服从正态分布

检验两个样本平均数差异是否显著，即检验 是否来源于同一个总体，或者说两样本代表的总体是否有显著差异

$$t(n_1+n_2-2)=\frac{\overline{X_1}-\overline{X_2}}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}(\frac{1}{n_1}+\frac{1}{n_2})}$$

配对样本T检验

前提：配对样本差独立

单样本*t*检验的扩展, 零假设：*μ=μ0*为真的条件下服从自由度为*n−1* 的t分布

$$t(n-1)=\frac{\overline{d}-u_0}{s_d/\sqrt{n}}$$

$$s_d=\sqrt{\frac{\sum_{i=1}^{n}(d_i-\overline{d})^2}{n-1}}$$

卡方分布与检验

卡方分布:

若随机变量X₁，X₂，..., Xn 独立同分布于N(0, 1) 则平方和服从卡方分布, n足够大时近似正态分布

$$\chi^2(n)=\sum_{i=1}^{n}X_i^2$$

$$E(\chi^2)=n,\ D(\chi^2)=2n,\ \chi^2(n_1)+\chi^2(n_2)\ \chi(n_1+n_2)[\text{两个卡方独立时}]$$

卡方检验:

卡方检验针对分类变量。常用在分类资料统计推断中：两个率或两个构成比比较的卡方检验；多个率或多个构成比比较的卡方检验以及分类资料的相关分析等。卡方检验就是统计样本的实际观测值与理论推断值之间的偏离程度，实际观测值与理论推断值之间的偏离程度就决定卡方值的大小，如果卡方值越大，二者偏差程度越大；反之，二者偏差越小；若两个值完全相等时，卡方值就为0，表明实际值与理论值完全符合

$$\chi^2=\sum\frac{(f_0-f_e)^2}{f_e}$$

两个分类变量是否独立最常用卡方检验，原假设是无关性假设，自由度=(行-1)*(列-1)

python模块—— *scipy.stats (chi2, chi2_contingency)*;

卡方检验还常用来进行特征选择，通过计算因变量与自变量的卡方值选择优于阈值的特征，或者直接选择k个最好特征。Python中是通过结合使用 *sklearn.feature_selection.SelectKBest*和*sklearn.feature_selection.chi2*实现的

F分布与检验(ANOVA)

F分布

$$X_1\ \chi^2(n_1),\ X_2\ \chi^2(n_2)[\text{独立}],\ F(n_1,n_2)=\frac{X_1/n_1}{X_2/n_2}$$

Z检验与T检验的不足

当研究中出现两个以上的平均数时，用Z检验和t检验会有以下一些不足:

1. 比较的组合次数增多
2. 降低可靠程度

对数据做得Z检验或t检验越多，我们更容易犯Ⅰ型错误。在一个检验中，α = 0.05，意味着有0.05的可能性犯Ⅰ型错误，即有1-α=0.95的概率不犯Ⅰ型错误。如果我们做两次检验，每次都为0.05的显著性水平，那么不犯Ⅰ型错误的概率就变为0.95×0.95 = 0.90。此时犯Ⅰ型错误的概率则为1-0.90 = 0.10，即至少犯一次Ⅰ型错误的概率翻了一倍。若做10次检验的话，至少犯一次Ⅰ型错误的概率将上升到0.40（1-0.952），而10次检验结论中都正确的概率只有60%。所以说采用Z检验或t检验随着均数个数的增加，其组合次数增多，从而降低了统计推论可靠性的概率，增大了犯错误的概率

3. 缺少综合或整体信息

采用Z检验或t检验都只提供了两个组的信息，而忽略了其余的综合信息。然而在许多情况下这些被忽视的信息可能对检验结果产生更大的影响力。同时在十次检验之后所得到只是零散的信息，并非从总体来分析几种不同条件的效果

方差分析可以避免以上不足

方差分析(ANOVA)

python模块—— *statsmodels.stats.anova*

方差分析的原理

所谓方差分析(*analysis of variance*)就是对多个平均数进行比较的一种统计方法，又称变异数分析，即ANOVA

总变异的分解： 总变异(SST) = 组间变异+组内变异 组间变异(SSA) = 实验条件 + 随机误差 组内变异(SSE) = 个体差异 + 实验误差 , 组内误差都是随机误差。

当差异主要来源于组内差异，即各样本均值并无显著差异时，各样本对目标值的影响没有显著差异；

当差异主要来源于组间差异，即各样本均值有显著差异时，各样本对目标值的影响有显著差异

方差分析前提条件

1. 各样本应该是相互独立的随机样本
2. 各样本来自正态分布总体
3. 各总体方差相等(方差齐性)

方差分析的用途**

分类型自变量对数值型因变量的影响

1. 两个或多个样本均数间的比较
2. 分析两个或多个因素间的交互作用

单因素方差分析

平方和	自由度	均方	F比
SSA	s-1	MSA = SSA/(s-1)	F = MSA/MSE
SSE	n-s	MSE = SSE/(n-s)	
SST	n-1		

双因素方差分析

单独效应 ——其他因素固定，某一因素不同水平之间均数的差别

交互效应 ——某因素的单独效应，随另一因素水平而变化，且不能用随机误差解释

假设因素A有r个水平（A1，A2.....，Ar），因素B有s个水平（B1，B2.....，Bs），每个AB交叉组(比如属于A1和B2的样本)进行t次独立试验，样本总数n

SST = SSA+SSE+SSB+SSAB

误差来源	平方和	自由度	均方	F比
因素A	SSA	r-1	MSA = SSA/(r-1)	FA= MSA/MSE
因素B	SSB	s-1	MSEB= SSB/(s-1)	FB = MSB/MSE
交互作用	SSAB	(r-1)(s-1)	MSAB = SSAB/{(r-1)(s-1)}	FAB = MSAB/MSE
误差	SSE	rs(t-1)	MSE = SSE/{rs(t-1)}	
综合	SST	rst-1		

参考文献

百度百科

维基百科

igoslly : https://www.cnblogs.com/igoslly/p/6782963.html

声明

因为是方便自己看的，所以笔记中省略了一部分我很熟悉的内容和公式，需要读者有一些数学或统计学基础

本笔记为个人整理，仅限学习使用，转载请标明作者和来源。码字不易，如果觉得不错，git上请点个***star***吧