.....?

[github: https://github.com/hfl15]

数学基础（2）~ 数理统计基础知识

出处：http://www.cnblogs.com/fanling999/p/6708458.html
参考：盛骤, 谢式千, 潘承毅. 概率论与数理统计, 第四版[M]. 高等教育出版社, 2008.

数理统计基础知识

_数理统计_是具有广泛应用的一个数学分支，它以概率论为理论基础，根据试验或观察得到的数据，来研究随机现象，对研究对象的客观规律性作出种种合理的估计和假设。

在实际工程中，我们对于一个总体进行研究，往往只能通过对总体的观察样本进行研究，基于样本的分布来研究总体的分布，数理统计为这样的过程提供和很好的支持。本文主要分为三个部分旨在对数理统计知识进行简要的回顾和总结，因此忽略了很多细节，如需要可以参考本文使用的教材，或其他相关书籍。

第一部分对抽样分布的内容进行了回顾总结，是后续章节的基础。根据大数定理，我们可以基于样本对总体的统计量进行合适的估计，统计量有样本均值、样本方差、样本标准差、样本k阶（原点）矩、样本k阶中心矩。使用统计量的分布（即抽样分布）对总体分布进行研究，总结了常用的三大分布即 χ^2，t分布，和F分布，主要关注分布的概率密度函数以及分为点。

第二部分和第三部分总结了统计推断的两大类问题，即估计问题和假设检验问题。

第二部分，参数估计，可分为点估计和区间估计。其中点估计有矩估计发和极大似然估计法。为了获知估计的可信程度，可使用区间估计法，其核心在于基于统计量的分布，以及分为点，确定参数估计区间。

第三部分，假设检验，是根据样本所提供的信息来考虑对假设作出接收或拒绝的决策过程。假设检验与区间估计类似，假设检验中有零假设和备择假设。我们总是假设在假设零假设正确的基础上去计算‘当零假设正确时被拒绝的概率’，这也被称为第一类错误发生的概率，并尽可能的减小这种错误发生的可能性，使得错误发生的概率很小，而小概率事件在一次试验中是几乎不可能发生的，因此对于一次观察，如果这样的错误发生了，我们就有理由怀疑零假设的正确性，从而做出拒绝零假设的决策，具体过程参考相应章节。最常用的假设检验方法有t检验。其中需要注意的问题还有样本容量的选取、原假设和备择假设的选取等。在实践中我们常喜欢使用p-value来衡量假设检验的显著程度，显著水平α相对应。最后，大多假设都基于分布已知的前提，这些也被称为参数化方法。然而实践中这不总是能获知，这个时候可以有两个解决方案（1）当样本容量充分大时，我们可以根据中心极限定理，使用正态分布对总体分布进行近似（2）使用非参数化方法，不需要基于分布已知的前提，不过其检验效果往往差于参数化方法，其中秩和检验就是这样的非参数化检验方法。因此在最后总结了分布拟合检验，对未知总体是否服从某一分布进行假设检验。

1.样本及抽样分布

本章主要介绍总体、随机样本及统计量等基本概念，介绍了几个常用的统计量和抽样分布。

1.1 随机样本

基本概念

- **总体：**实验全部可能的观察值
- **个体：**每个观察值被成为个体
- **容量：**总体中所包含的个体的个数
- **有限总体：**容量有限
- **无限总体：**容量无限（有些有限总体的容量很大，可以认为是无限总体，例如考察全国正在使用的某种灯泡的寿命）

公告

昵称：.....?
园龄：5年10个月
粉丝：51
关注：3
+加关注

< 2019年6月

日	一	二	三	四
26	27	28	29	30
2	3	4	5	6
9	10	11	12	13
16	17	18	19	20
23	24	25	26	27
30	1	2	3	4

搜索

随笔分类(88)

ASP.NET(3)
c/c++(5)
leetcode(20)
Windows开发相关(17)
读书(4)
环境(4)
机器学习与数据挖掘(6)
数学基础(2)

- **样本：**在数理统计中，人们都是通过从总体中抽取一部分个体，根据获得的数据来对总体分布作出推断的，被抽出的部分个体叫做总体的一个样本。
- **抽样：**放回抽样和不放回抽样。对于有限总体，采用放回抽样可以得到**简单随机样本**，但放回抽样使用起来不方便，因此当个体总数N比要得到的样本的容量n大得多时，可将不放回抽样当作放回抽样来处理。对于无限总体，因抽取一个个体不影响它的分布，因此总是使用不放回抽样。
- **简单随机样本：**_在相同条件下对总体X进行n次重复的、独立的观察_，n次观察结果依次表示为$X_1,X_2,...,X_n$，可认为他们相互独立并都是与总体X具有相同分布的随机变量。

重要定义

定义：设X是具有分布函数F的随机变量，若 $X_1,X_2,...,X_n$ 是具有同一分布函数F的、相互独立的随机变量，则称 $X_1,X_2,...,X_n$ 为从分布函数F（或总体F、或总体X）得到的**容量为n的简单随机样本**，简称**样本**，它们的观察值 $x_1,x_2,...,x_n$ 称为**样本值**，又称为X的n个**独立的观察值**。

将样本表示成一个随机向量（$X_1,X_2,...,X_n$），对应的样本值为 （$x_1,x_2,...,x_n$）。
若（$x_1,x_2,...,x_n$）和（$y_1,y_2,...,y_n$）都是相应于样本（$X_1,X_2,...,X_n$）的样本值，一般来说它们是不相同的。
由定义（$X_1,X_2,...,X_n$）的分布函数为：

$$F^*(x_1,x_2,\ldots,x_n)=\prod_{i=1}^{n}F(x_i)$$

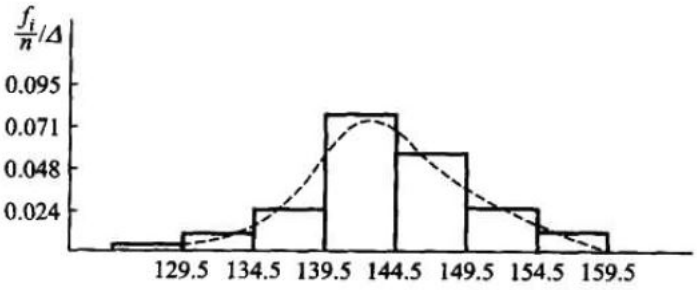
若X具有概率密度，则 （$X_1,X_2,...,X_n$）的概率密度为：

$$f^*(x_1,x_2,\ldots,x_n)=\prod_{i=1}^{n}f(x_i)$$

1.2 直方图和箱线图

图表是进行数据分析的有效工具，这里给出两个常用的基本统计图：

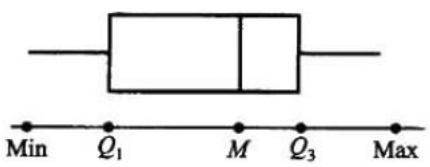
频率直方图：（1）将可能的结果分成几个区间，即横坐标的分段，统计每个分段的频率并作图（1）小矩形面积=数据落在该区间内的频率。



箱线图：
（1）图中各个值的意思： Min、Max分别表示样本的最大值和最小值，M表示样本中位数（又称第二四分位点，对应0.5分位数$x_{0.5}$），Q1表示第一个四分位点（即0.25分位数，$x_{0.25}$），Q3表示第三四分位点（对应0.75分位数$x_{0.75}$）。

（2）几个概念：

- 中心位置：中位数M所在位置就是数据集的中心；
- 离散程度：全部数据都落在[Min,Max]之内，[Min,Q1],[Q1,M],[M,Q3],[Q3,Max]区间内的数据个数约各占1/4。区间较短时，表示落在区间的点比较集中，反之较为分散；
- 对称性：若中位数位于箱子的中间位置，则数据分布较为对称（下图中Min离M的距离较Max离M的距离大，表示数据分布左倾斜，反之可称为右倾斜）。



疑似异常值： 在数据集中不寻常的大于或小于数据集中的其他数据的值。（箱线图中小于Min和大于Max部分的离群点）

算法学习(26)

综合(1)

最新评论

1. Re:HOOK API（四）——止

您好，能发一份源码吗？103qq.com

2. Re:HOOK API （一）——础+一个鼠标钩子实例

认真学习中， 顺便说一下代码UTTONDOWN写错了写成双

3. Re:使用R进行相关性分析

你好，请问如果数据集中既有类别变量（factor或characte做相关性矩阵？直接将类别变值型变量后用cor函数意义岂变了？比如：area这个变量，3.....

4. Re:使用R进行相关性分析

@我是酱油妹呀我建议你看官一个tutorial专门讲如何画图，细。...

5. Re:使用R进行相关性分析

你好，我用PerformanceAna了相关图，但是字体、大小、合杂志社发文章的要求，不知改？可以教教我吗，或有什么推荐吗？几乎逛遍各种贴吧论相应.....

--我

推荐排行榜

异常来源：（1）数据的测量、记录或输入计算机时的误差；（2）数据来自不同的总体；（3）数据是正确的，但它只体现小概率事件。（面对异常需进一步分析原因，如果无法对异常进行处理，选择模型时应考虑对异常或离群点具有鲁棒性较好的模型）

1.3 抽样分布

样本是进行统计推断的依据。在应用时，往往不是直接使用样本本身，而是针对不同的问题构造样本的适当函数，利用这些样本的函数进行统计推断。

1.3.1 几个常用的统计量

定义：设$X_1,X_2,...,X_n$是来自总体X的一个样本，$g(X_1,X_2,...,X_n)$ 是 $X_1,X_2,...,X_n$的函数，g中不含未知参数，则称 $g(X_1,X_2,...,X_n)$ 是一**统计量**。

> 因为$X_1,X_2,...,X_n$是随机变量，而$g(X_1,X_2,...,X_n)$是随机变量的函数，因此统计量是一个随机变量。相应的设 $x_1,x_2,...,x_n$ 为样本值，那么$g(x_1,x_2,...,x_n)$为$g(X_1,X_2,...,X_n)$的观察值。

- **样本平均值：** Sample Mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- **样本方差：**（分母为n-1是为了保证无偏估计） Sample Variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2)$$

- **样本标准差：** Sample Standard deviation

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- **样本k阶（原点）矩：**

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, k = 1,2,\ldots;$$

- **样本k阶中心矩：**

$$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k, k = 2,\ldots;$$

将样本值$x_1,x_2,...,x_n$带入上面的统计量表达式则可以得到对应的观察值。

1.3.2 经验分布函数

总体分布函数F(X)的统计量称为**经验分布函数**。

构造方法如下：设$X_1,X_2,...,X_n$是总体F的一个样本，用S(x)，-inf < x < inf 表示 $X_1,X_2,...,X_n$ 中不大于 x 的随机变量的个数，则经验分布函数定义如下：

$$F_n(x) = \frac{1}{n}S(x), -\infty < x < \infty$$

一般，设$x_1,x_2,...,x_n$是总体F的一个容量为n的样本值，先将$x_1,x_2,...,x_n$ 按自小到大的次序排列，并重新编号。设为 $x_{(1)} <= x_{(2)} <=...<=x_{(n)}$，则经验分布函数的观察值为：

$$F_n(x) = \begin{cases} 0, & if\, x < x_1 \\ \frac{k}{n} & if\, x_k \leqslant x < x_{k+1}, k = 1,2,\ldots,n-1 \\ 1, & if\, x \geq x_n \end{cases}$$

对于经验分布函数，格里汶科（Glivenko）在1933年已经证明，当n趋于无穷时，经验分布函数一致收敛于分布函数F(x)。因此当n充分大时，经验分布函数的任一观察值与总体分布函数F(x)只有微小的差别，实际上可以当作总体分布F(x)来只用。

1.3.3 几个常用统计量的分布

在使用统计量进行统计推断时常需要知道它的分布，然而这在实际中是困难的。因此我们经常使用统计量的分布来研究总体分布，而统计量的分布称为**抽样分布**。

下面给出三个来自正态分布的抽样分布，即统计学中的三大分布，重点给出定义、概率密度函数图、分位点

（1）χ^2（卡方分布）

1. 面试经典算法题集锦——《r》小结(13)

2. HOOK API（三）—— HOOK 序的 MessageBox(5)

3. HOOK API （一）——HOOK +一个鼠标钩子实例(4)

4. HOOK API（四）—— 进

5. 使用R进行相关性分析(2)

定义： 设$X_1,X_2,...,X_n$是来自标准正态总体N(0,1)的样本，则称统计量

$$\chi^2 = X_1^2 + X_2^2 + \ldots + X_n^2$$

服从自由度为n的χ^2分布，记为$\chi^2 \sim \chi^2(n)$
（自由度是指独立变量的个数）

性质：

- 可列可加性

$$\chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2)\ ,\ \chi_1^2和\chi_2^2相互独立性$$

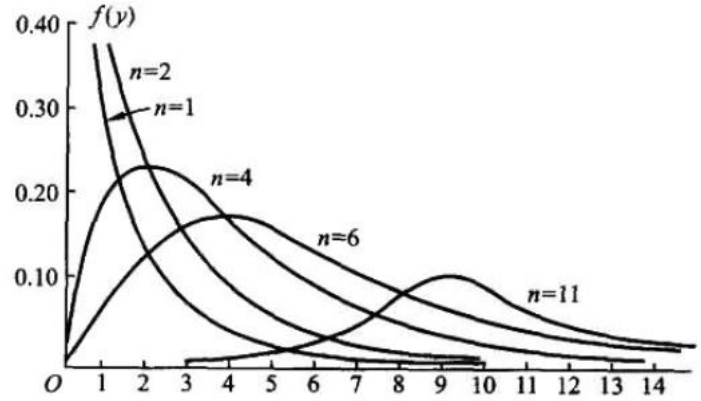
- 数学期望和方差（根据定义很好证明）

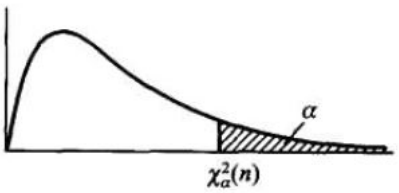
$$E(\chi^2) = n, D(\chi^2) = 2n$$

- 上分位点（参考图形，计算查表）

$$P(\chi^2 > \chi_\alpha^2(n)) = \int_{\chi_\alpha^2(n)}^{\infty} f(y)dy = \alpha$$

概率密度在n不同取值下的图形；上分为点示意图。





（2） t分布

定义： 设 $X \sim N(0,1), Y \sim \chi^2(n)$ 且 X，Y相互独立，则称随机变量

$$t = \frac{X}{\sqrt{Y/n}}$$

服从自由度为n的t分布，记为t~t(n)。

上分位点

$$P(t > t_{\alpha(n)}) = \int_{t_\alpha(n)}^{\infty} h(t)dt = \alpha$$

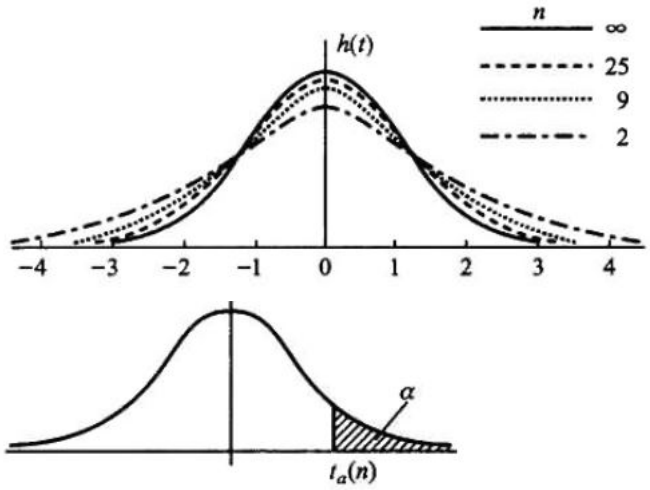
对称性：

$$t_{1-\alpha}(n) = -t_\alpha(n)$$

当n>45时，可用正态近似：

$$t_\alpha(n) \approx z_\alpha$$

t分布的概率密度图；上分为点图示。



（3） F分布

定义： 设 $U \sim \chi^2(n_1), V \sim \chi^2(n_2)$ 且U,V相互独立，则称随机变量

$$F = \frac{U/n_1}{V/n_2}$$

服从自由度为(n_1, n_2)的F分布，记为 F~F(n_1, n_2)。

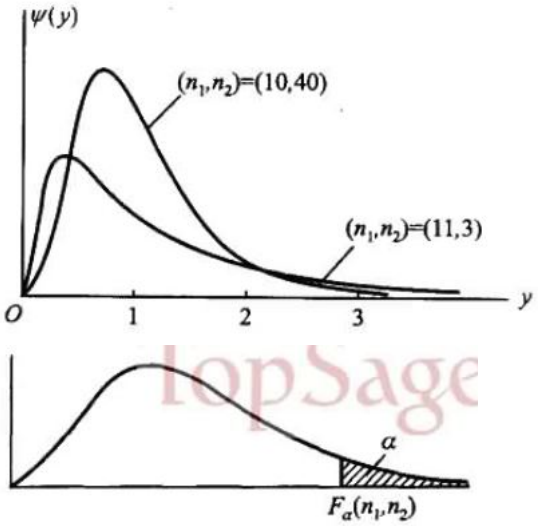
由定义可知：1/F ~ F(n_2, n_1)

上分位点

$$P(F > F_\alpha(n_1, n_2)) = \int_{F_\alpha(n_1,n_2)}^{\infty} \varphi(y)dy = \alpha$$

$$F_{1-\alpha}(n_1, n_2) = \frac{1}{F_\alpha(n_1, n_2)}$$

F分布的概率密度图；上分位点示意图



注意：在分为点中 $0 < \alpha < 1$

1.3.4 正态总体的样本均值和样本方差的分布

（1）设X_1,X_2,...,X_n是来自总体X（不管服从什么分布，只要它的均值和方差存在）的样本，并且有：

$$E(X) = \mu, D(X) = \sigma^2$$

则有：

$$E(\bar{X}) = \mu, D(\bar{X}) = \sigma^2/n$$

（2）设总体 $X \sim N(\mu, \sigma^2)$， $X_1,X_2,...,X_n$ 是来自总体X的样本，则有：

- $\bar{X} \sim N(\mu, \sigma^2/n)$
- $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$
- $\bar{X}, S^2$ 相互独立
- $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$
- 两个正态总体 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$
  - $\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$
  - 当 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 查看参考书

2.参数估计

参数估计问题可以分为：点估计和区间估计。点估计是适当的选择一个统计量作为未知参数的估计，若已取得一样本，将样本值带入估计量，得到估计量的值，以估计量的值作为未知参数的值。点估计不能反应估计的精度，因此引入了区间估计，置信区间是一个随机区间，其具有高的预先给定的概率覆盖未知参数。

2.1 点估计

定义：设总体X的分布函数的形式已知，但它的一个或多个参数未知，借助于总体X的一个样本来估计未知参数的值的问题称为参数的**点估计问题**。下面主要总结两种常用的点估计方法，即：矩估计法和最大似然估计法。

点估计的一般提法：设总体X的分布函数 $F(x;\theta)$ 的形式为已知 θ 是待估参数。$X_1,X_2,...,X_n$是X的一个样本，$x_1,x_2,...,x_n$是相应的一个样本值。点估计问题就是要构造一个适当的统计量$\hat{\theta}(X_1, X_2, \ldots, X_n)$，用它的观察值$\hat{\theta}(x_1, x_2, \ldots, x_n)$作为未知参数$\theta$的近似值。我们称$\hat{\theta}(X_1, X_2, \ldots, X_n)$为$\theta$的**估计量**，称$\hat{\theta}(x_1, x_2, \ldots, x_n)$为$\theta$的**估计值**。

（1）矩估计法

1. 基于样本矩依概率收敛于总体矩构造估计量，即：

$$A_l = \frac{1}{n}\sum_{i=1}^{n} X_i^l \xrightarrow{P} \mu_l, l = 1, 2, \ldots, k$$

2. 根据样本矩估计总体矩得关于未知k个参数的方程组，即：

$$\begin{cases} \mu_1 = & \mu_1(\theta_1, \theta_2, \ldots, \theta_k) \\ \mu_2 = & \mu_2(\theta_1, \theta_2, \ldots, \theta_k) \\ \cdots & \\ \mu_k = & \mu_k(\theta_1, \theta_2, \ldots, \theta_k) \end{cases}$$

3. 根据k个方程组解出未知参数，即：

$$\begin{cases} \theta_1 = & \theta_1(\mu_1, \mu_2, \ldots, \mu_k) \\ \theta_2 = & \theta_2(\mu_1, \mu_2, \ldots, \mu_k) \\ \cdots\cdots & \\ \theta_k = & \theta_k(\mu_1, \mu_2, \ldots, \mu_k) \end{cases}$$

使用样本矩代替总体矩得到：
\left{

$$\theta_1 = \theta_1(A_1, A_2, \ldots, A_k)\ \theta_2 = \theta_2(A_1, A_2, \ldots, A_k) \ldots\ldots \theta_k = \theta_k(A_1, A_2, \ldots, A_k)$$

\right.

（2）最大似然估计法

1. 结合联合概率和条件概率的计算，可得样本$X_1,X_2,...,X_n$观察到值$x_1,x_2,...,x_n$的概率如下（称为样本的**似然函数**）：

$$L(\theta) = L(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} p(x_i, \theta), \theta \in \Theta$$

2. 原理：小概率事件在一次试验中几乎不可能发生，因此可以认为当前观察到的样本值$x_1,x_2,...,x_n$发生的概率较大，即$ L(\theta)较大，我们不会考虑那些不能使当前样本值出现的那些\theta \in \Theta作为未知参数的估计，而是应该考虑那些使得 L(\theta)

较大的参数作为估计。＊＊＊＊由费希尔($R.A.Fisher$)引进的最大似然估计法，就是固定样本观察值$x < sub > 1 < /sub >, x\< s$.,$x < sub > n < /sub >$ ，在
theta的可能范围内挑选使得似然函数L(x_{1},x_{2},...,x_{n};\theta)达到最大值的参数值\hat_{\theta}$作为估计值，
即：****

$$L(x_1, x_2, \ldots, x_n; \hat{\theta}) = max_{\theta \in \Theta} L(x_1, x_2, \ldots, x_n; \theta)$$

$\hat{\theta}(x_1, x_2, \ldots, x_n)$称为参数$\theta$的**最大似然估计值**，而相应的$\hat{\theta}(X_1, X_2, \ldots, X_n)$称为参数$\theta$的**最大似然估计量**。

3. $lnL(\theta)$ 和 $L(\theta)$在同一θ处取到极值，因此可以对似然函数取对数后求解（取对数的操作可以将乘法转换为加法，计算上要更为简单），即对k个方程解以下微分方程得到未知参数的估计：

$$\frac{\partial}{\partial \theta_i} lnL = 0, i = 1, 2, \ldots, k$$

注意，对于连续型随机变量，似然函数可取（使用概率密度函数）：

$$L(\theta) = L(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i, \theta), \theta \in \Theta$$

2.2 区间估计

区间估计是确定未知参数的一个取值范围，并给出未知参数落入这个范围的一个概率估计即可信程度。
定义： 假总体X的分布函数$F(x;\theta)$含有一个未知参数$\theta, \theta \in \Theta$($\Theta$是可能取值的范围)，对于给定值$\alpha(0 < \alpha < 1)$，若由来自X的样本$X_1,X_2,...,X_n$确定的两个统计量$\underline{\theta} = \underline{\theta}(X_1, X_2, \ldots, X_n)$ 和 $\bar{\theta} = \bar{\theta}(X_1, X_2, \ldots, X_n)$ ($\underline{\theta} < \bar{\theta}$)，对于任意 $\theta \in \Theta$满足

$$P(\underline{\theta}(X_1, X_2, \ldots, X_n) < \theta < \bar{\theta}(X_1, X_2, \ldots, X_n)) \geqslant 1 - \alpha$$

则称随机区间($\underline{\theta} < \bar{\theta}$)是$\theta$置信水平为$1 - \alpha$ 的 **置信区间**，$1 - \alpha$称为**置信水平**，$\underline{\theta}$为**置信下限**，$\bar{\theta}$为**置信上限**。

一般步骤：

1. 寻找一个样本$X_1,X_2,...,X_n$和θ的函数$W = W(X_1, X_2, \ldots, X_n; \theta)$，使得W的分布不依赖于$\theta$以及其他未知参数，称具有这种性质的函数W为**枢轴量**。（可以从上一章的抽样分布入手进行构造）
2. 对于给定的置信水平$1 - \alpha$，定出两个常数a，b使得 $P(a < W(X_1, X_2, \ldots, X_n; \theta) < b) = 1 - \alpha$。若能从 $a < W(X_1, X_2, \ldots, X_n; \theta) < b$ 得到与之等价的θ的不等式$\underline{\theta} < \theta < \bar{\theta}$，那么($\underline{\theta} < \bar{\theta}$)是$\theta$置信水平为$1 - \alpha$ 的 **置信区间**。（根据上一步构造的枢轴量所服从分布的上分为点进行确定）

注意：枢轴量$W = W(X_1, X_2, \ldots, X_n; \theta)$的构造，通常可以从$\theta$的点估计着手考虑。常用的正态总体的参数的置信区间可以用上述步骤推得。

一个例子：
问题：设总体$X \sim N(\mu, \sigma^2)$，σ^2为已知，μ为未知，设$X_1, X_2, \ldots, X_n$是来自X的样本，求μ的置信水平为$1 - \alpha$的置信区间。
解答：
我们知道$\bar{X}$是μ的无偏估计，且有：

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

$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$服从标准正态分布不依赖于任何未知参数。按标准正太分布的上α分为点的定义可得（如下图所示）：

$$P(|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}| < z_{\alpha/2}) = 1 - \alpha$$

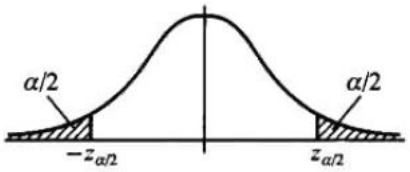
概率表示图中无阴影，中间部分。由此解得：

$$P(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}) = 1 - \alpha$$

由此可以得到μ的一个置信水平为$1 - \alpha$的置信区间：$(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2})$
最后只需要带入变量并查表就可以得到确切的区间。

值得注意的是，满足要求的置信区间不止一个，两个端点的面积加起来为α则满足要求，但其中$\alpha/2$分为点形成的置信区间最短，因此精度最好，所以被选为置信区间（具体可参考课本P163）。

标准正态分布的分为点：



下面给出常用的区间估计，其不同在于枢轴量的构建，因此只给出各种情况下数轴量的表示以及服从的分布

2.3 正态总体均值与方差的区间估计

1. 单个总体$X \sim N(\mu, \sigma^2)$

(1) 均值μ的置信区间

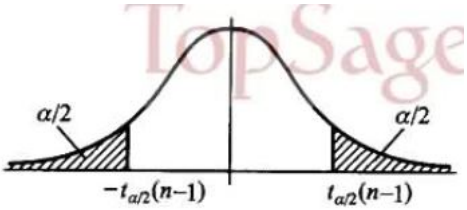
1.1 σ^2 已知

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

如上文的例子

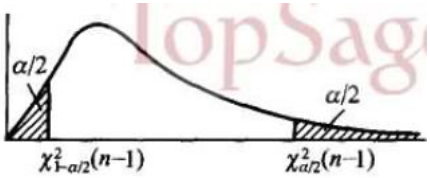
1.2 σ^2 未知

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



(2) 方差σ^2的置信区间

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



2. 两个总体$X \sim N(\mu_1, \sigma_2^2), Y \sim N(\mu_2, \sigma_2^2)$

(1) 两个总体均值差$\mu_1 - \mu_2$的置信区间

$$\bar{X}-\bar{Y} \sim N(\mu_1-\mu_2, \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2})$$

(2) 两个总体方差比σ_1^2/σ_2^2的置信区间

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

2.4 （0-1）分布参数的区间估计

有中心极限定理，当n充分大时有：

$$\frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}} = \frac{n\bar{X}-np}{\sqrt{np(1-p)}} \sim N(0,1)$$

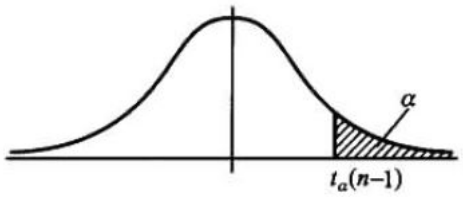
2.5 单侧置信区间

单侧致信区间是确定参数的上限或则下限，只需要根据给定的置信度确定上分为点或下分为点即可，如下面两图所示，其求解过程与双侧区间类似。

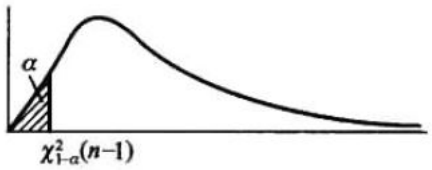
$$P(\theta > \underline{\theta}) \geqslant 1-\alpha$$

$$P(\theta < \bar{\theta}) \geqslant 1-\alpha$$

t分布的上α分为点：



卡方分布的下α分位点，可以根据性质求得（参考上一章）：



2.6 估计量的评选标准

用不同的估计方法求出的估计量可能不相同，原则上任何统计量都可以作为未知参数的估计。至于哪一个更好，有以下3个常用的评判标准，即无偏性、有效性和相合性。

1. 无偏性

若估计量$\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$的数学期望存在，且对于任意$\theta \in \Theta$有$E(\hat{\theta}) = \theta$，这称$\hat{\theta}$ 为 θ 的无偏估计。

估计量相对于真值来说总会存在一定的误差，偏大或者偏小，无偏性是要求反复对估计量使用多次，其均值可以逼近真值，即要求系统误差$E(\hat{\theta}) - \theta$为0.

2. 有效性

有效性是对估计量离散程度的一个考量，对于两个无偏估计量，方差小的要更优。

3. 相合性

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1$$

估计量要依概率收敛于真值，这是估计量的基本要求，如果估计量不能满足相合性，那么不论样本容量n取多么大，都不能得到参数的准确估计，是不可取的。

2.7 基于截尾样本的最大似然估计

很多时候由于各方面因素，比如时间和经济的因素，我们不能获取到完全样本。因此就会存在截断抽样，可分为**定时结尾样本**和**定数结尾样本**。以研究灯泡的寿命为例：**定时结尾样本**是给定一个观察终止的时间点，观察在这个时间点内有多少灯泡失效；**定数结尾样本**是给定常数m，当失效的灯泡数量达到m时，实验结束，得到一个样本。 对于这类问题，关键在于确定似然函数。

3.假设检验

有关总体分布的未知参数或未知分布形式的种种论断叫统计假设，人们根据样本所提供的信息对所考虑的假设作出接受或拒绝的决策。假设检验就是作出这一决策的过程。

3.1 假设检验

处理参数的假设检验问题的步骤如下：

1. 根据实际问题的要求，提出原假设H_0及备择假设H_1
2. 给定显著水平α以及样本容量n
3. 确定检验统计量以及拒绝域的形式
4. 按P{当H_0为真拒绝H_0}$\leqslant \alpha$求出拒绝域
5. 取样，根据样本观察值作出决策，是接受H_0还是拒绝H_0

示例

在显著水平α下，检验假设：

$$H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$$

H_0称为**原假设**或**零假设**

H_1称为**备择假设**

假设检验的过程是：我们认为H_0假设是正确的，并尝试根据样本统计量对均值的真值进行估计，这个时候均值的无偏估计$\bar{X}$应该与μ_0非常接近，即$|\bar{X} - \mu_0|$不会过分的大，如果很不幸对于某一样本值$|\bar{x} - \mu_0|$过大，又基于小概率事件在一次实验中几乎不可能发生，然而现在发生了，那么我们就有理由怀疑H_0假设的正确性。通常来说，我们会给定一个阈值k以控制是否接受H_0假设的决策。

另一方面，$|\bar{x} - \mu_0|$的大小与$\frac{|\bar{X}-\mu_0|}{\sigma/\sqrt{n}}$的大小是正相关的，而后者作为统计量更容易计算，因此我们往往会从某一统计量入手去做决策。既然是决策，就就有可能发生错误，即当H_0为真时，我们仍然有可能将其拒绝，这也被称为假设检验中的第一类错误，我们希望尽可能减小这类错误发生的概率，

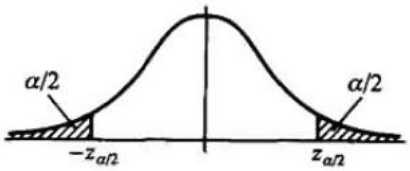
P{当H_0为真拒绝H_0} = $P_{\mu_0}(|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}| \geq k) = \alpha$

解释：H_0为真，但其样本均值$\bar{X}$与给定值的偏离程度超出了阈值k，这个时候我们将会做出拒绝H_0。然而！！！H_0是真的，因此我们犯了第一类错误，而我们希望折中错误发生的概率很小，即α很小，往往取0.1,0.05,0.01,0.005等值。

H_0为真时，$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$，由标准正态分布分为点的定义，可以得到$|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}| \geq k = z_{\alpha/2}$，如下图：

对于任一样本值，计算$|z| = |\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}|$，如果|z|大于$z_{\alpha/2}$，小概率事件发生了，那么我们有理由怀疑原假设的正确性，因此拒绝原假设，否则我们没有足够的理由拒绝原假设。

标准正态分布的分为点，我们希望阴影部分的面积尽可能小（这是犯第一类错误的概率，也是拒绝域）：



常用的正态总体均值、方差的假设检验

根据中心极限定理，当样本容量很大时，很多分布都可以近似到正态分布进行处理。假设检验有双边检验、单边检验（左边检验和右边检验）。

t检验是实践中最常用到的假设检验，因为实践中往往很难获知方差的情况。对于单个正态总体，可以使用t检验均值的是否产生显著变化。对于两个正态总体，分两种情况（1）输入的是两组不同环境下的观察值，那么使用一般的t检验（2）输入是两组相同条件下的成对的（对比实验的）观察值，可以使用成对数据的t检验。（参考下面的表格）

对于单一实验样本可以采用t检验，对于成对的观察值可以采用成对的t检验。

表 8—1　正态总体均值、方差的检验法(显著性水平为 α)

	原假设 H_0	检验统计量	备择假设 H_1	拒绝域
1	$\mu\leqslant\mu_0$ $\mu\geqslant\mu_0$ $\mu=\mu_0$ (σ^2 已知)	$Z=\dfrac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$	$\mu>\mu_0$ $\mu<\mu_0$ $\mu\neq\mu_0$	$z\geqslant z_\alpha$ $z\leqslant -z_\alpha$ $\lvert z\rvert\geqslant z_{\alpha/2}$
2	$\mu\leqslant\mu_0$ $\mu\geqslant\mu_0$ $\mu=\mu_0$ (σ^2 未知)	$t=\dfrac{\overline{X}-\mu_0}{S/\sqrt{n}}$	$\mu>\mu_0$ $\mu<\mu_0$ $\mu\neq\mu_0$	$t\geqslant t_\alpha(n-1)$ $t\leqslant -t_\alpha(n-1)$ $\lvert t\rvert\geqslant t_{\alpha-2}(n-1)$
3	$\mu_1-\mu_2\leqslant\delta$ $\mu_1-\mu_2\geqslant\delta$ $\mu_1-\mu_2=\delta$ (σ_1^2,σ_2^2 已知)	$Z=\dfrac{\overline{X}-\overline{Y}-\delta}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}$	$\mu_1-\mu_2>\delta$ $\mu_1-\mu_2<\delta$ $\mu_1-\mu_2\neq\delta$	$z\geqslant z_\alpha$ $z\leqslant -z_\alpha$ $\lvert z\rvert\geqslant z_{\alpha/2}$
4	$\mu_1-\mu_2\leqslant\delta$ $\mu_1-\mu_2\geqslant\delta$ $\mu_1-\mu_2=\delta$ ($\sigma_1^2=\sigma_2^2=\sigma^2$ 未知)	$t=\dfrac{\overline{X}-\overline{Y}-\delta}{S_w\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$ $S_w^2=\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$	$\mu_1-\mu_2>\delta$ $\mu_1-\mu_2<\delta$ $\mu_1-\mu_2\neq\delta$	$t\geqslant t_\alpha(n_1+n_2-2)$ $t\leqslant -t_\alpha(n_1+n_2-2)$ $\lvert t\rvert\geqslant t_{\alpha/2}(n_1+n_2-2)$
5	$\sigma^2\leqslant\sigma_0^2$ $\sigma^2\geqslant\sigma_0^2$ $\sigma^2=\sigma_0^2$ (μ 未知)	$\chi^2=\dfrac{(n-1)S^2}{\sigma_0^2}$	$\sigma^2>\sigma_0^2$ $\sigma^2<\sigma_0^2$ $\sigma^2\neq\sigma_0^2$	$\chi^2\geqslant\chi_\alpha^2(n-1)$ $\chi^2\leqslant\chi_{1-\alpha}^2(n-1)$ $\chi^2\geqslant\chi_{\alpha/2}^2(n-1)$或 $\chi^2\leqslant\chi_{1-\alpha/2}^2(n-1)$
6	$\sigma_1^2\leqslant\sigma_2^2$ $\sigma_1^2\geqslant\sigma_2^2$ $\sigma_1^2=\sigma_2^2$ (μ_1,μ_2 未知)	$F=\dfrac{S_1^2}{S_2^2}$	$\sigma_1^2>\sigma_2^2$ $\sigma_1^2<\sigma_2^2$ $\sigma_1^2\neq\sigma_2^2$	$F\geqslant F_\alpha(n_1-1,n_2-1)$ $F\leqslant F_{1-\alpha}(n_1-1,n_2-1)$ $F\geqslant F_{\alpha/2}(n_1-1,n_2-1)$或 $F\leqslant F_{1-\alpha/2}(n_1-1,n_2-1)$
7	$\mu_D\leqslant 0$ $\mu_D\geqslant 0$ $\mu_D=0$ (成对数据)	$t=\dfrac{\overline{D}-0}{S_D/\sqrt{n}}$	$\mu_D>0$ $\mu_D<0$ $\mu_D\neq 0$	$t\geqslant t_\alpha(n-1)$ $t\leqslant -t_\alpha(n-1)$ $\lvert t\rvert\geqslant t_{\alpha/2}(n-1)$

3.2 假设检验的其他关键内容：

1. 置信区间与假设检验之间的关系

实际上置信区间是对某一参数的区间估计，这一区间对应着相应的假设检验中的接受域，$1-\alpha$置信水平的置信区间，对应着α显著水平的假设检验的接受域。我们在进行假设检验（显著性检验）时更关注拒绝域。

2. 假设检验中的两类错误

第I类错误是假设检验中显式控制的错误，又称为"弃真"，第II类错误称为"存伪"。

假设检验的两类错误

真实情况（未知）	所作决策	
	接受 H_0	拒绝 H_0
H_0 为真	正确	犯第Ⅰ类错误
H_0 不真	犯第Ⅱ类错误	正确

3. 样本容量的选取

在假设检验中，总是根据问题的要求，预先给出显著性水平以控制犯第I类错误的概率，而犯第II类错误的概率则依赖于样本容量的选择。一些实际问题中，我们除了希望控制犯第I类错误的概率外，往往还希望控制犯第II类错误的概率。这里可以通过OC曲线来进行研究。

4. 假设检验问题的p值法

定义：假设检验问题的p值（probability value）是由检验统计量的样本观察值得出的原假设可被拒绝的最小显著水平。

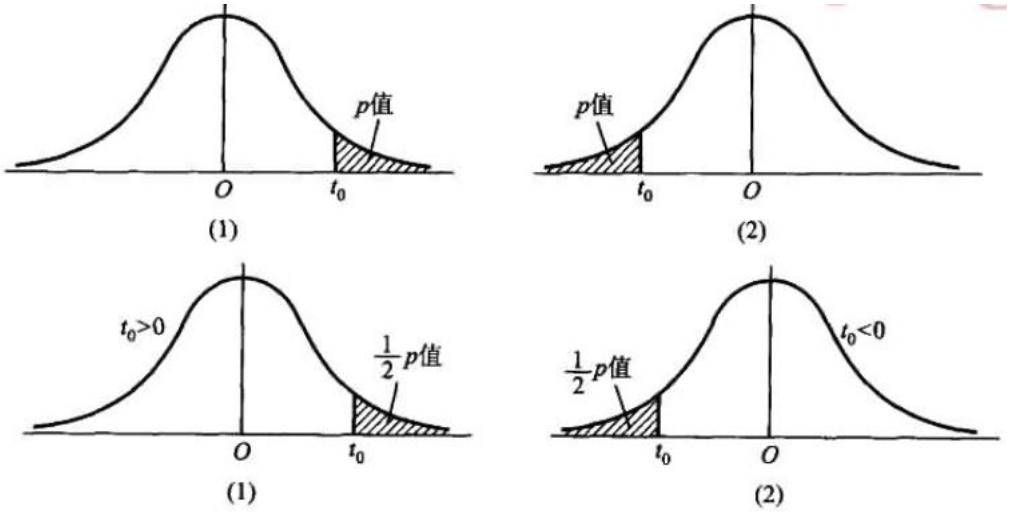
按p值的定义，对于任意显著性水平α，就有：
(1)若p值<=α，则在显著性水平α下拒绝H_0
(2)若p值>α，则在显著性水平α下接受H_0

在现代计算机统计软件中，一般都给出检验问题的p值。

p值表示反对原假设H_0的依据的强度，p值越小，反对H_0的依据越强、越充分。
一般，p值的讨论可以分为以下几种情况：

- 若p<=0.01，称推断拒绝H_0的依据很强或称检验是**高度显著的**；
- 若0.01<p<=0.05称推断拒绝H_0的依据很强或称检验是**显著的**；
- 若0.05<p<0.1称推断拒绝H_0的理由是弱的，检验是**不显著的**；
- 若p>0.1一般来说**没有理由**拒绝H_0。

t单边检验和双边检验的p value：



5. 原假设和备择假设的选择

在进行显著性检验时，犯第I类错误的概率是由我们控制的。α取得小，保证了当H_0为真时错误地拒绝H_0的可能性很小。这意味着H_0是受到保护的，也表明H_0、H_1的地位是不对等的。于是，在一对对立假设中，选哪一个作为H_0需要小心。
一般情况下，选择H_0、H_1使得两类错误中后果严重的错误成为第一类错误，这是选择H_0、H_1的一个原则。 比如考虑某种药品是否为真时，应该将'药品为假'作为H_0，第一类错误就是'药是假的但被拒绝了'，也就是说'药是真的'，这个存在很大的危险性，不过现在我们将其作为H_0假设，我们可以控制减小犯这种严重错误的概率。
如果两类错误中，没有一类错误的后果严重更需要避免时，常常取H_0为维持现状，即取H_0为'无效益'，'无改进'，'无价值'等，这样会比较保守一些。
在实际问题中，情况比较复杂，如何选取H_0，H_1，只能在实践中积累经验，根据实际情况去判断。

3.4 秩和检验

显著性检验的方法可以分为参数统计方法和非参数统计方法。

(1)参数统计方法：总体分布类型已知，用样本指标对总体参数进行推断或假设检验的方法。

(2)非参数统计方法：不用考虑总体分布是否已知，不比较总体参数，只比较总体分布的位置是否相同的统计的方法。

前面提及的统计检验方法，比如t检验，均属于参数统计方法，需要提前知道总体分布的形式。一般情况下，当样本容量足够大时，基于中心极限定理，可使用正态分布（高斯分布）作为近似。

而秩和检验是典型的非参数化统计方法，不需要知道总体分布的形式，不过值得注意的是检验需要满足'独立性'是前提。

3.3 分布拟合检验

实际问题中，总体的分布往往不总是可以被获取到的，这时需要根据样本检验关于分布的假设。课本中主要介绍了χ^2拟合检验法，它可以用来检验总体是否具有某一个指定的分布或属于某一个分布族。此外还介绍了专门用于检验分布是否为正态的"偏度、峰度检验法"。

(1)单个分布的χ^2拟合检验法

(2)分布族的χ^2拟合检验法

(3)偏度、峰度检验

随机变量的偏度和峰度是指X的标准化变量$[X-E(X)]/\sqrt{D(X)}$的三阶矩和四阶矩：

$$\nu_1 = E[(\frac{X-E(X)}{\sqrt{D(X)}})^3] = \frac{E[(X-E(X))^3])}{(D(X))^{3/2}}$$

$$\nu_2 = E[(\frac{X-E(X)}{\sqrt{D(X)}})^4] = \frac{E[(X-E(X))^4])}{(D(X))^2}$$

当随机变量X服从正太分布时$\nu_1 = 0$且$\nu_2 = 3$。

https://github.com/hfl15

分类： 数学基础

标签： 机器学习， 数据挖掘， 概率论， 数学

好文要顶 关注我 收藏该文

.....?
关注 - 3
粉丝 - 51
+加关注

0 0

« 上一篇：数学基础（1）~ 概率论基础知识
» 下一篇：面试中的概率题

posted @ 2017-04-16 20:18 .....? 阅读(1029) 评论(0) 编辑 收藏

刷新评论 刷新页面 返回顶部

注册用户登录后才能发表评论，请 登录 或 注册， 访问 网站首页。

【推荐】超50万C++/C#源码：大型实时仿真组态图形源码
【前端】SpreadJS表格控件，可嵌入系统开发的在线Excel
【推荐】程序员问答平台，解决您开发中遇到的技术难题

相关博文：
· 概率论与数理统计常用英文词汇对照
· 概率论与数理统计常用英文词汇对照
· 概率论与数理统计-ch7-参数估计
· 数学基础-概率论05（统计推断-分布拟合检验）
· 【数理统计学习】统计假设检验

最新新闻：
· 开发、推广"数据精灵"外挂干扰微信运营 法院一审判赔500万
· 华为与俄最大电信公司签约 将在俄罗斯开发5G网络
· 看上了人家的工程师？传苹果正收购自动驾驶创企
· 穿越宇宙的电波
· 惠普CEO：Intel处理器缺货将使得AMD处理器份额持续提升
» 更多新闻...

Copyright ©2019 .....?