韩非囚秦
——独善其身者，难成大事也。

导航
博客园
首页
新随笔
联系
订阅
管理

< 2021年6月 >

日	一	二	三	四	五	六
30	31	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	1	2	3
4	5	6	7	8	9	10

统计学基本原理

1.随机事件

确定性现象：在一定条件下必然发生的现象称为确定性现象；特征：条件完全决定结果

随机现象：在一定条件下可能出现也可能不出现的现象称为随机现象；特征：条件不能完全决定结果。

随机现象是通过随机试验来研究的。具有以下三个特征的试验称为随机试验：

(1)可以在相同的条件下重复进行；

(2)每次试验的可能结果不止一个，并且能事先明确试验的所有可能结果；

(3)进行一次实验之前不能确定哪一个结果会出现。

样本空间和样本点：定义随机试验E的所有可能的结果组成的集合称为E的样本空间，记为Ω。样本空间的元素，即试验E每一个结果，称为样

随机事件：随机试验E的样本空间的子集称为E的随机事件。

对于抛筛子试验：它的样本空间是{1,2,3,4,5,6},每一个元素就是样本点，"大于3的概率"是随机事件。因此有$\Omega \geq A \geq \omega_i$

2.随机事件的关系

事件的交:事件A与事件B同时发生，则称这样一个事件为交或者积，记为$A \cap B$或者AB;

事件的并:事件A与事件B至少有一个发生，也即A和B的所有样本点构成的集合，称为并，记为$A \cup B$;

事件的包含: 事件A包含事件B，记为$A \supset B$;

事件的相等:事件A与事件B相等，记为$A = B$

事件的互斥:如果事件A与事件B的交集为空$(AB = \phi)$，则称A和B互斥;

事件的差:事件A发生而B不发生，记为$A - B$;

事件的对立如果事件A和B有且仅有一个发生，且他们的并集是整个集合$(A \cup B = \Omega$，且$A \cap B = \phi)$

随机事件的独立性是各种数学模型的基本前提假设

2.随机事件的规律性--概率

频率的定义：在相同的条件下进行了n次试验，在这n次试验中，事件A发生的次数n_A称为事件A发生的频数，比值$\frac{nA}{n}$称为事件A发生的频率

频率不是概率

随机事件A的概率：一般地，在大量重复试验中，如果事件A发生的频率m/n会稳定在某个常数p附件，那么这个常数p就叫做事件A的概率，

概率的性质:

(1)对于任意事件A，有:$0 \leq P(A) \leq 1$

(2)对于必然事件A和不可能事件B，有$P(\text{必然事件}) = 1$，$P(\text{不可能事件}) = 0$

(3)对于两两互斥的可数个事件$A_1, A_2, \ldots, A_n$，有$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n) = P(A)$，称P

(4)$P(\overline{A}) = 1 - P(A)$

(5)$A \subset B$，则$P(A) \geq P(B)$

事件的独立性与条件概率:

设A，B为两事件，且$P(A) > 0$，称$P(B|A) = \frac{P(AB)}{P(A)}$为事件A发生的条件下事件B发生的条件概率;

设A，B为两事件，且满足公式$P(AB) = P(A)P(B)$，则称A与B事件独立。

设$A_1, A_2, \ldots, A_n$是n个事件，如果其两两互斥，则有$P(A_1 A_2 \ldots A_n) = P(A_1)P(A_2) \ldots P(A_n)$

五大公式(极其重要):

(1)加法公式:

$$P(AUB) = P(A) + P(B) - P(AB)$$

$$P(AUBUC) = P(A) + P(B \cup C) - P\left((A \cap B)U(A \cap C)\right) = P(A) + P(B) + P(C) - P(BC) - P(AB) -$$

(2)减法公式:

$$P(A - B) = P(A) - P(AB)$$

(3)乘法公式:

当$P(A) > 0$时，有$P(AB) = P(A)P(B|A)$

--heroisuseless
3. Re:前端(七)：ES6一些新特性
map不是es6的吧,ie9都支持map方法
--普通男孩

阅读排行榜
1. 决策树之ID3算法(22684)
2. 一、python简单爬取静态网页(9617)
3. Django(三)：HttpRequest和HttpResponse(5590)
4. python(五)：元类与抽象基类(4282)
5. tensorflow(一)：图片处理(4192)

评论排行榜
1. 决策树之ID3算法(2)
2. 前端(七)：ES6一些新特性(1)

推荐排行榜
1. Django(四)：model(1)
2. 决策树之ID3算法(1)
3. KNN算法(1)
4. python(五)：元类与抽象基类(1)

$当P(A_1 A_2 ... A_n)>0时，有P(A_1 A_2 ... A_n) = P(A_1)P(A_2|A_1) ... P(A_n|A_1 A_2 ... A_{n-1})$

(4)全概率公式[先验概率公式]:

设$B_1, B_2, \ldots, B_n$满足 $\cup_{i=1}^{n} B_i = \Omega$，$B_i B_j = \phi(i \neq j)$且$P(B_i) > 0$，则对任意事件A有:

$$P(A) = \sum_{i=1}^{n} P(B_i)P(A|B_i)$$

(5)贝叶斯公式[后验概率公式]:

设$B_1, B_2, \ldots, B_n$满足 $\cup_{i=1}^{n} B_i = \Omega$，$B_i B_j = \phi(i \neq j)$且$P(B_i) > 0$，对于$P(A) > 0$，有:

$$P(B_j|A) = \frac{P(b_j)P(A|B_j)}{\sum_{i=1}^{n} P(B_i)P(A|B_i)}$$

二、随机变量及其概率分布

1.随机变量

定义:在样本空间$\Omega上的实值函数X=X(\omega)，\omega \in \Omega，称X(\omega)为随机变量，记为X$

2.分布函数

定义:对于任意实数x，记函数$F(x) = P\{X \leq x\}, -\infty < x < +\infty$，称$F(x)$为随机变量$X$的分布函数，$F(x)$的值等于随机变量$X$在区

显然地，F(x)具有下列性质:

(1) $0 \leq F(x) \leq 1$

(2)$F(x)$是单调非减函数，即当$x_1 < x_2$，$F(x_1) \leq F(x_2)$

(3)$F(x)$是右连续的，即$F(x+0) = F(x)$

(4)对任意的$x_1 < x_2$，有$P\{x_1 < X < x_2\} = F(x_2) - F(x_1)$

(5)对任意的x, $P\{X = x\} = F(x) - F(x-0)$

3.离散型随机变量X的概率分布

设离散型随机变量X的可能取值是$x_1, x_2, \ldots, x_n$，X取各可能的值得概率为 $P\{X = x_k\} = P_k, k = 1, 2, ..$称上式为离散型随机变量X的

<table>
<tr><td rowspan="10">离散型随机变量及其分布</td><td rowspan="2">随机变量及其分布列</td><td>概念</td><td colspan="2">随着试验结果变化而变化的量叫做随机变量，所有取值可以一一列出的随机叫做离散型随机变量。</td></tr>
<tr><td>分布列</td><td colspan="2">离散型随机变量的所有取值及取值的概率列成的表格  。</td></tr>
<tr><td>性质</td><td colspan="2">（1） $p_i \geq 0 (i = 1,2,\cdots, n)$；（2） $p_1 + p_2 + \cdots + p_n = 1$。</td></tr>
<tr><td rowspan="3">事件的独立性</td><td rowspan="2">条件概率</td><td colspan="2">概念：事件 A 发生的条件下，事件 B 发生的概率， $P(B|A) = \frac{P(AB)}{P(A)}$。</td></tr>
<tr><td colspan="2">性质：$0 \leqslant P(B|A) \leqslant 1$， B, C 互斥， $P(B \cup C|A) = P(B|A) + P(C|A)$。</td></tr>
<tr><td>独立事件</td><td colspan="2">事件 A 与事件 B 满足 $P(AB) = P(A)P(B)$，事件 A 与事件 B 相互独立。</td></tr>
<tr><td>n 次独立重复试验</td><td colspan="2">每次试验中事件 A 发生的概率为 p，在 n 次独立重复试验中，事件 A 恰好发生 k 次的概率为 $P(X = k) = C_n^k p^k (1-p)^{n-k}, (k = 0,1,2,\cdots, n)$。</td></tr>
<tr><td rowspan="4">典型分布</td><td>超几何分布</td><td colspan="2">$P(X = k) = \frac{C_M^k C_{N-M}^{n-k}}{C_N^n}$， $k = 0,1,2,\cdots, m$，其中 $m = \min\{M, n\}$，且 $n \leqslant N$，且 $n \leq N, M \leq N, n, M, N \in \mathbf{N}^*$。</td></tr>
<tr><td>二项分布</td><td colspan="2">分布列为： $P(X = k) = C_n^k p^k (1-p)^{n-k}, (k = 0,1,2,\cdots, n)$， $X \sim B(n, p)$。数学期望 $EX = np$、方差 $DX = np(1-p)$【$n=1$时为两点分布】</td></tr>
<tr><td>正态分布</td><td colspan="2">$\varphi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}$ 图象称为正态密度曲线，随机变量 X 满足 $P(a < X \leqslant b) = \int_a^b \varphi(x)dx$，则称 X 的分布为正态分布。正态密度曲线的特点。</td></tr>
<tr><td rowspan="2">数字特征</td><td>数学期望</td><td>$EX = x_1 p_1 + x_2 p_2 + \cdots + x_i p_i + \cdots + x_n p_n$</td><td>$E(aX+b) = aEX + b$</td></tr>
<tr><td>方差和标准差</td><td>方差：$DX = \sum_{i=1}^{n} (x_i - EX)^2 p_i$，标准差：$\sigma X = \sqrt{DX}$</td><td>$D(aX+b) = a^2 DX$</td></tr>
</table>

4.连续型随机变量及其概率分布

如果对随机变量X的分布函数$F(x)$，存在一个非负可积函数$f(x)$，使得对任意函数x，都有$F(x) = \int_{-\infty}^{x} f(t)d(t), -\infty < x < +\infty$，

概率密度函数f(x)的性质:

(1)$f(x) \geq 0$

(2)$\int_{-\infty}^{+\infty} f(x)dx = 1$

(3)对任意实数$x_1 < x_2$，有$P\{x_1 < X \leq x_2\} = \int_{x_1}^{x_2} f(t)dt$

(4)在$f(x)$的连续点处有$F'(x) = f(x)$，如果X是连续型随机变量，则显然有$P\{x_1 < X \leq x_2\} = P\{x_1 \leq X < x_2\} = P\{x_1$

三.随机变量的数字特征

1.数学期望:

离散型随机变量的数学期望:

已知随机变量X的概率分布为$P\{X = x_k\} = P_k, k = 1, 2, \ldots,$ 则$E(X) = \sum_{k=1}^{+\infty} x_k P_k$

连续型随机变量的数学期望:

已知随机变量X的概率密度为$f(x)$，其概率分布为$\int_{-\infty}^{x} f(t)dt$，则$E(X) = \int_{-\infty}^{+\infty} x f(x)dx$

数学期望的性质:

设X是随机变量，C是常数，则有:$E(CX) = CE(X)$

设X和Y是任意两个随机变量，则有:$E(X \pm Y) = E(X) \pm E(Y)$
设随机变量X和Y相互独立，则有:$E(XY) = E(X)E(Y)$

2.方差:

设X是随机变量，如果数学期望$E\{[X - E(x)]^2\}$存在，则称为X的方差，记作$D(X)$，即$D(X) = E\{[X - E(X)]^2\}$。称$\sqrt{D(x)}$

方差计算公式: $D(X) = E(X^2) - [E(X)]^2$

分　布	分布律或概率密度	EX	DX
0-1分布 $X \sim B(1,p)$	$P\{X=k\}=p^k(1-p)^{1-k}$ $k=0,1$	p	$p(1-p)$
二项分布 $X \sim B(n,p)$	$P\{X=k\}=C_n^k p^k(1-p)^{n-k}$ $k=0,1,\cdots,n$	np	$np(1-p)$
泊松分布 $X \sim P(\lambda)$	$P\{X=k\}=\frac{\lambda^k}{k!}e^{-\lambda}$ $k=0,1,2,\cdots$	λ	λ
几何分布 $X \sim G(p)$	$P\{X=k\}=(1-p)^{k-1}p$ $k=1,2,\cdots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
均匀分布 $X \sim U[a,b]$	$f(x)=\begin{cases}\frac{1}{b-a}, & a\le x\le b,\\ 0, & \text{其它.}\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
指数分布 $X \sim E(\lambda)$	$f(x)=\begin{cases}\lambda e^{-\lambda x}, & x>0,\\ 0, & x\le 0.\end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
正态分布 $X \sim N(\mu,\sigma^2)$	$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2

3.矩、协方差、相关系数

矩:

原点矩：设X是随机变量，如果$E(X)^2$，k=1,2,...存在，则称之为X的k阶原点矩

中心距：设X是随机变量，如果$E\{[X - E(X)]^k/\}$存在，则称之为X的k阶中心距

协方差:

对于随机变量X和Y，如果$E\{[X - E(X)][Y - E(Y)]\}$存在，则称之为X和Y的协方差，记作$cov(X,Y)$即:

$$cov(X,Y) = E\{[X - E(X)][Y - E(Y)]\}$$

显然地，$X - E(X)$和$Y - E(Y)$是两个标准差的向量表示形式(标准差是内积)，它的物理意义是反映了两个向量的夹角和其模之间的

相关系数:

对于随机变量X和Y，如果$D(X)D(Y) \neq 0$，则称$\frac{cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$为X和Y的相关系数，记为$\rho_{XY}$

它们之间的关系及推导公式详见：https://blog.csdn.net/dcrmg/article/details/52416832

四、数理统计的基本概念

1.基本概念

总体：数理统计中所研究对象的某项数量指标X的全体称为总体。

样本：如果$X_1, X_2, \ldots, X_n$相互独立且都与总体X同分布，则称$X_1, X_2, \ldots, X_n$为来自总体的简单随机样本，n为样本容量，样本的具体

统计量：样本$X_1, X_2, \ldots, X_n$的不含未知参数的函数$T = T(X_1, X_2, \ldots, Xn)$称为统计量。

	总体参数	样本统计量
定义	反映总体数量特征的指标	反映样本数量特征的指标
符号	总体容量 N 总体均值 $\overline{X}$ μ 总体成数 π 总体方差 σ^2 总体标准差 σ	样本容量 n 样本均值 $\overline{x}$ 样本成数 p 样本方差 s^2 样本标准差 s

样本数字特征：设$X_1, X_2, \ldots, X_n$是来自总体X的样本，则称:

(1)样本均值:

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

(2)样本方差:

$S^2 = \frac{1}{n-1}\sum_{i-1}^{n}(X_i - \overline{X})^2$，样本标准差开根号即可;

(3)样本k阶原点矩:

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, k = 1, 2, A_1 = \overline{X}$$

(4)样本k阶中心距:

$$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k, k = 1, 2, B_2 = \frac{n-1}{n}S^2 \neq S^2$$

样本数据特征的性质:

(1)如果总体X具有数学期望$E(X) = \mu$，则:

$$E(\overline{X}) = E(X) = \mu$$

备注:意思是，如果总体X的数学期望存在，那么它的数学期望就等于样本的均值，即样本均值是总体均值的无偏估计量

(2)如果总体X具有方差$D(X) = \sigma^2$，则:

$$D(\overline{X}) = E(S^2) = D(X)/n = \sigma^2/n$$

备注:意思是，如果总体X的方差存在，那么它的方差除以样本量就等于样本的方差，并且样本方差是总体方差的无偏估计量

(3)平均偏差:$\frac{\sqrt{|X-u|}}{N}$

(4)离散系数:标准差与其相应的均值之比，表示为百分数。用于比较两组数据离散程度[变异程度]的大小

五、参数[抽样]估计

1.理论基础:

抽样估计就是从总体中抽样，计算样本均值、方差、成数等参数，以此梯段总体参数的过程。

抽样推断的理论基础:

1.大数定律:频率以及大量测量值的算术平均值具有稳定性，不受个别测量值的影响。

2.大量随机变量和的分布近似于正态分布。这里衍生了独立同分布的各种极限定理。

2.参数估计方法

点估计:

用样本$X_1, X_2, \ldots, X_n$构造的统计量$\hat{\theta}(X_1, X_2, \ldots, X_n)$来估计未知参数$\theta$称为点估计，统计量$\hat{\theta}(X_1, X_2, \ldots, X_n)$称为估计量

无偏估计量:

设$\hat{\theta}$是θ的估计量，如果$E(\hat{\theta}) = \theta$，则称$\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$是未知参数$\theta$的无偏估计量。

一致估计量:

设$\hat{\theta}(X_1, X_2, \ldots, X_n)$是$\theta$的估计值，如果$\hat{\theta}$依概率收敛于$\theta$，则称$\hat{\theta}(X_1, X_2, \ldots, X_n)$是$\theta$的一致估计量。

**证明样本均值是总体数学期望的无偏估计量:

已知:$E(\overline{X}) = E(X) = \mu$

推导:$E(X) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$

**证明样本方差是总体方差的无偏估计量:

已知:$D(\overline{X}) = E(S^2)/n = D(X)/n = \sigma^2/n$

推导:

$E(S^2) = \frac{1}{n-1}E\{\sum_{i=1}^{n}[(X_i - \mu) - (\overline{X} - \mu)]^2\} = \frac{1}{n-1}E\{\sum_{i=1}^{n}[(X_i - \mu)^2 - 2(X_i - \mu)(\overline{X} - \mu) + (\overline{X} - \mu)^2]\} = \frac{1}{n-1}E[\sum$

抽样平均误差:$\mu_{\overline{x}} = \frac{\sigma(X)}{\sqrt{N}}$

区间估计:在一定的概率保证程度下，选定一个区间δ，再根据样本指标数值和δ去估计总体指标数值所在的可能范围的一种统计推断方法。

(1)置信区间:设$theta$是总体X的未知参数，$X_1, X_2, \ldots, X_n$是来自总体X的样本，对于给定的$\alpha(0 < \alpha < 1)$，如果两个统计量满足

$$P\theta_1 < \theta < \theta_2 = 1 - \alpha$$

则称随机区间(θ_1, θ_2)为参数θ的置信水平(或置信度)为$1 - \alpha$的置信区间(或区间估计)，简称为]\theta的1-\alpha的置信区间，\\theta_1 和 \theta_2分别称为置信下限和置信上限$

(2)整理:

估计区间的上下限:$\Delta_{\overline{x}}$，相当于下面第二张表第一行的$\frac{\sigma}{\sqrt{n}}Z_{\frac{\alpha}{2}}$

置信区间:$[\overline{x} \pm \Delta_{\overline{x}}]$

置信度$F(t) = P(|\overline{x} - \overline{X}| \leq t\mu_{\overline{x}})$

t称为概率度，它与置信度存在分布上的转换关系，如下图所示。这里的$\mu_{\overline{x}}$就相当于下面第二张表第一行的$\frac{\sigma}{\sqrt{n}}$，也即总体标准差。

常用概率度与置信度对照表

概率度t	误差范围△	置信度
1.00	$1\mu_{\overline{x}}$	68.27%
1.64	$1.64\ \mu_{\overline{x}}$	90%
1.96	$1.96\ \mu_{\overline{x}}$	95%
2.00	$2\ \mu_{\overline{x}}$	95.45%
3.00	$3\mu_{\overline{x}}$	99.73%

(3)区间估计的求解过程：

以下面表中第一行的前提条件为例。

根据样本资料计算$\bar{x}$和$\frac{\sigma}{\sqrt{n}}$；

根据给定的置信度查正态分布表计算概率度

根据上述公式计算估计区间。

备注：就是根据大数定律，大量样本和的分布接近正态分布，并在正态分布上继续构造各种统计量来计算给定置信度下的均值和方差的置信区

正态总体均值和方差置信区间一览表（置信水平为$1-\alpha$）

	待估参数	其它参数	W所服从的分布	置信区间	单侧置信区间
一个正态总体	均值μ	σ^2 已知	$Z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$	$\left(\overline{X}\pm\frac{\sigma}{\sqrt{n}}Z_{\frac{\alpha}{2}}\right)$	$\overline{\mu}=\overline{X}+\frac{\sigma}{\sqrt{n}}Z_{\alpha}$，$\underline{\mu}=\overline{X}-\frac{\sigma}{\sqrt{n}}Z_{\alpha}$
		σ^2 未知	$T=\frac{\overline{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$	$\left(\overline{X}\pm\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)\right)$	$\overline{\mu}=\overline{X}+\frac{S}{\sqrt{n}}t_{\alpha}(n-1)$，$\underline{\mu}=\overline{X}-\frac{S}{\sqrt{n}}t_{\alpha}(n-1)$
	方差σ^2	μ 未知	$\chi^2=\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$	$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)},\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right)$	$\overline{\sigma^2}=\frac{(n-1)S^2}{\chi^2_{\alpha}(n-1)}$，$\underline{\sigma^2}=\frac{(n-1)S^2}{\chi^2_{1-\alpha}(n-1)}$
两个正态总体	均值差$\mu_1-\mu_2$	σ_1^2,σ_2^2 已知	$Z=\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1)$	$\left(\overline{X}-\overline{Y}\pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$	$\overline{\mu_1-\mu_2}=\overline{X}-\overline{Y}+Z_{\alpha}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$ $\underline{\mu_1-\mu_2}=\overline{X}-\overline{Y}-Z_{\alpha}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
		$\sigma_1^2=\sigma_2^2=\sigma^2$ 未知	$T=\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2)$	$\left(\overline{X}-\overline{Y}\pm t_{\alpha/2}(n_1+n_2-2)S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$	$\overline{\mu_1-\mu_2}=\overline{X}-\overline{Y}+t_{\alpha}(n_1+n_2-2)S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ $\underline{\mu_1-\mu_2}=\overline{X}-\overline{Y}-t_{\alpha}(n_1+n_2-2)S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
	方差σ_1^2/σ_2^2	μ_1,μ_2 未知	$F=\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2}\sim F(n_1-1,n_2-1)$	$\left(\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha/2}(n_1-1,n_2-1)},\frac{S_1^2}{S_2^2}\frac{1}{F_{1-\alpha/2}(n_1-1,n_2-1)}\right)$	$\overline{\frac{\sigma_1^2}{\sigma_2^2}}=\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha}(n_1-1,n_2-1)}$ $\underline{\frac{\sigma_1^2}{\sigma_2^2}}=\frac{S_1^2}{S_2^2}\frac{1}{F_{1-\alpha}(n_1-1,n_2-1)}$

3.常用统计抽样分布和正态总体的抽样分布

卡方分布:

设随机变量$X_1,X_2,\ldots,X_n$相互独立且服从标准正态分布N(0，1)，则称随机变量$\chi^2=X_1^2+X_2^2+\ldots+X_n^2$服从自由度为n的卡方

性质:

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

设$\chi_1^2\sim\chi^2(n_1),\chi_2^2\ sim\chi^2(n_2)$，且$\chi_1^2$和$\chi_2^2$相互独立，则$\chi_1^2+\chi_2^2\sim\chi^2(n_1+n_2)$。

t分布:

设随机变量X和Y相互独立，且$X\sim N(0,1),Y\sim\chi^2(n)$，则称随机变量$T=\frac{X}{\sqrt{Y/n}}$服从自由度为n的t分布，记作$T\sim t(n)$。

性质:

t分布的概率密度是偶函数，和正态分布的概率密度函数非常相似，当n充分大时，t分布近似标准正态分布

F分布:

设随机变量X和Y相互独立，且$X\sim\chi^2(n_1),Y\sim\chi^2(n_2)$，则称随机变量$F=\frac{X/n_1}{Y/n_2}$服从自由度为$(n_1,n_2)$的F分布，记作$F\sim F($

性质: 它的导数也是F分布

统计三剑客的作用:

显然地，可以对均值和方差构造新的统计量，使其符合符合上述分布，从而进行区间估计及后面的显著性检验。

正态分布一般用于检验大样本量下的连续型数据的分布情况。

卡方分布用于分类变量的卡方检验。F分布多用于方差齐性检验。t分布用于小样本时的总体均值的检验。

六、假设检验

假设检验依据的统计原理是：小概率事件在一次实验中是不会发生的，又称小概率原理。

假设检验的两类错误：第一类错误，拒绝实际为真；第二类错误，接收实际为假。

显著性水平：在假设检验中允许犯第一类错误的概率，记为$\alpha(0<\alpha<1)$,则α称为显著性水平，它表现了对假设H_0的控制程度，一般α取

显著性检验：只控制第一类错误概率α的统计检验，称为显著性检验。

显著性检验的一般步骤：

1)根据问题要求提出原假设H_0

2)给出显著性水平α

3)确定检验统计量及拒绝形式

4)按犯第一类错误的概率等于α求出拒绝域W

5)根据样本值计算检验统计量T的观测值，当$t \in W$时，拒绝原假设H_0，否则，接收原假设H_0。

假设检验和区间估计的区别：

假设检验和区间估计过程相反，几乎可以看作是逆运算。

区间估计在已知的总体参数和样本参数的情况下，去估计总体的均值或方差的置信区间。在上表第一行中，假设知道了样本均值$\bar{x}$，样本
推测总体均值的置信区间就是上表第一行的置信区间。

同样地，假设检验在已知的总体参数和样本参数的情况下，去估计样本的均值或方差的置信区间。在上表第一行中，在给定的显著性水平
因为有$F(t) = P(|\bar{x} - \mu| < t * z_{\alpha/2})$

两者无非是和μ的计算而已。假设检验的表和上表一致。

p值和z值：

这里需要总结一下比较混乱的检验方式，以z检验为例。z检验的前提是总体方差已知。

$\alpha = 0.05$ 则计算它对应的置信区间[-1.96, 1.96]()，以下有三种计算方法来确定拒绝或接受原假设。

1.直接计算样本均值的估计区间，看抽取的样本是否落在估计区间内:

$$\mu - \sigma/\sqrt(n) * 1.96 < \bar{x} < \mu + \sigma/\sqrt(n) * 1.96$$

2.构造统计量，计算样本均值的概率度，概率度t是否落在置信区间内:

$z = \frac{\bar{x}-\mu}{\sqrt{\sigma/n}}$，看它是否落在[-1.96, 1.96]的置信区间内

3.计算了z值的概率度，继续计算p值，看它是否小于α:

$z = \frac{\bar{x}-\mu}{\sqrt{\sigma/n}}$，得到它的概率度，求它的双侧概率密度值，假设是z=2.15（p值是0.03），于是继续计算它的p值:

$$p = 2[1 - \phi(2.15)] = 2 * (1 - 0.98437) = 0.031$$

显然地，当$p > \alpha = 0.01$，拒绝原假设；当$p < \alpha = 0.05$，接受原假设

七、样本均值和方差检验的场景

均值检验：适用于均值是否存在差别的问题，反应的是集中趋势。

单样本均值检验：测试某批产品是否正常，或者某个部件是否正常，以及评价风险是否在可控范围内等。视总体方差已知和未知分为z检

双样本均值检验：测试两个总体的均值是否有差别。api: stats.ttest_ind和ttest_ind_from_stats

配对样本t检验：同一样本在某一条件影响的前后是否有差异。比如化肥与小麦产量，培训前后差异等。思路：两条数据相减得到一列数

方差检验：适用方差是否存在差别的问题，反应的离中趋势。

这里要说明因素及其水平。假如收入是目标变量，它受学历的影响。那么学历是一个因素，学历的等级是水平。试验的目的是查看不同

单因素方差分析：连续变量是否受某分类变量不同水平的影响。

多因素方差分析：已经过渡为一般线性模型，连续变量是否受某些分类变量的影响，以及分量变量对连续变量的影响是否受到别的分类变

分类: 机器学习

标签: 数据统计

一只火眼金睛的男猴
关注 - 0
粉丝 - 23
+加关注

« 上一篇： django基础一：web、wsgi、mvc、mtv
» 下一篇： 一、神经网络基础

posted on 2018-03-25 14:42 一只火眼金睛的男猴 阅读(2892) 评论(0) 编辑 收藏 举报